A Bayesian dynamic approach to modelling flow through a traffic network.

Thesis

UNRESTRICTED

A Bayesian Dynamic Approach to Modelling Flow through a Traffic Network

A thesis submitted to The Open University at Milton Keynes in the subject of statistics for the degree of doctor of philosophy

by

Benjamin John Wright BSc. MSc.

April 2005

DATE OF SUBMISSION 01 JULY 2005
DATE OF AWARD 10 MAY 2006

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I would like to thank my supervisor, Catriona Queen, for her invaluable insight, patience and above all hard work in helping me put this thesis together. Thanks are also due to all the members of the statistics department of the Open University for providing such a good research atmosphere and stimulating seminars in the time I have been here.

I would also like to thank Kent County Council, the original suppliers of the data used in this thesis.

Perhaps the warmest thanks (and sympathy) should be offered to everybody who suffered me bringing this work up in general conversation- in particular my parents and Claire.

Abstract

Hourly traffic flows through complicated motorway junctions form time series with a hierarchical element. Multiregression Dynamic Models provide a Bayesian framework for forecasting the time series while incorporating this hierarchical structure. Calculation of forecasts is computationally inexpensive and the model is designed in order to have readily interpretable parameters wherever possible. Expert intervention is straightforward in this system for periods of unusual activity, whether they were anticipated or not. Fundamental change to the road layout and behaviour can also be incorporated into the model without untenable complication. In this thesis work is centred on applying the methodology to one particular junction, including finding techniques to aid in using the methodology in this application. The theory of Multiregression Dynamic Models is furthered in order to do this- specifically through modelling certain parameter constraints imposed by the nature of the problem and the production of prior covariances for quantities modelled. The model was found to perform competitively in comparison to common other approaches, and the hierarchical structure offered significant advantages when expert intervention was applied.

Chapter 1 – Introduction

This chapter introduces the problem the thesis focuses on. The problem is outlined and existing approaches to similar problems discussed.

1.1 The problem

The accurate forecasting of road traffic can provide many benefits to those that are responsible for the maintenance and design of road networks, which can in turn provide benefits to those that use them. Good models can be used to respond to events, for example by identifying that an accident has occurred or by re-routing traffic to avoid congestion. They can also be used to forecast the consequences of future planned events such as roads being closed or built.

There are many aspects of a road network that can be modelled, including: traffic flow, traffic speed, journey time, route choice and state models of different traffic conditions. A forecaster will choose a model that is most suited for the purpose at hand. Models of such a network should therefore attempt to address one particular aspect or related group of aspects about the network.

In this thesis the aim is to provide forecasts for traffic flow counts through counting points situated at various places in the network. The model presented can easily incorporate information from outside the network (such as experts' estimation of the effects of future events) and still produce parameters that are readily understandable and have 'real-world' meaning. Such models can be used for real-time monitoring of the traffic network, to analyse current traffic patterns and to inform future decisions regarding the network.

The key goals are competitive performance of the model, ease of interpretation of the quantities in the model and the ability to incorporate exogenous information. There are secondary goals that are also desirable- a methodology that does not rely on extensive calculations so that it can be used in a real-time system and one that can produce information that may aid in the incorporation of exogenous data.

1.2 The data

Kent County Council installed traffic counting systems around the M25/A2/A282 motorway junction in Dartford. These systems count the number of vehicles that pass them each hour. Each counting system is identified by a number, and together they form a multivariate time series. The geographical layout of these counting points is illustrated in the figure 1.2.1.

The counting systems are distributed in such a way that traffic will flow into the network, through a number of counting points, then out of the network. These counting points can be categorised as one of three types. The first type consists of locations where all the vehicles passing them are entering the network for the first time (such as nodes 160 and 167). The second type consists of points where all the traffic passing them has passed a previous counting point in the network (such as nodes 161 and 168). The third type, predictably, consists of nodes where the traffic flowing past is a mixture of both types (there are no such nodes in this network). A model that incorporates the relationships between points will need to model each type of counting point in a slightly different way. Determining the category of each counting point is covered in section 2.2.

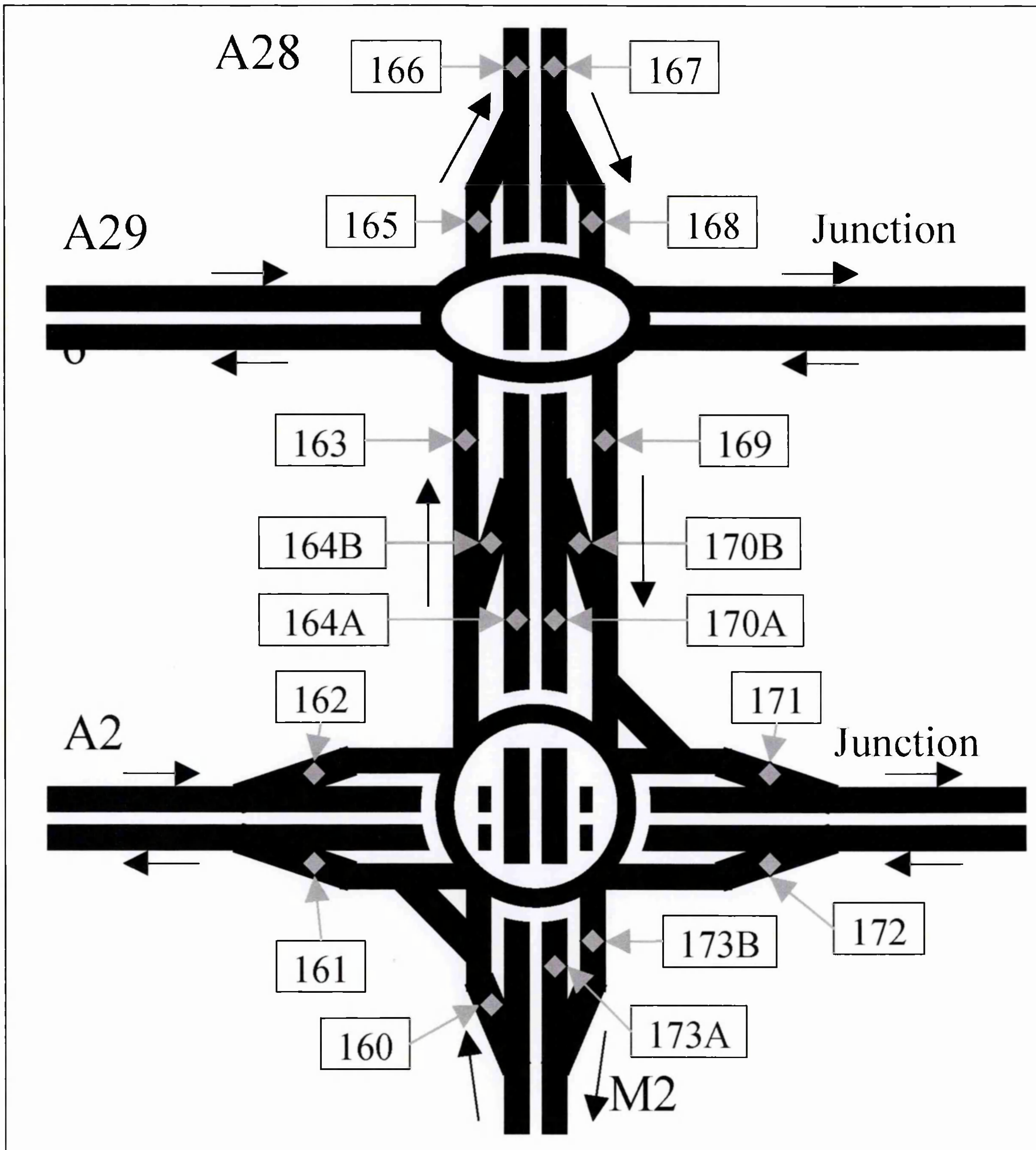

Figure 1.2.1 – M25, A2, A282 junction layout.

As the counts are hourly this will have a smoothing effect on short-term events- so minor events may not be observed in the collected data. Further, the time taken to traverse the network is short- during normal conditions it will only take a handful of minutes- compared to the counting interval of an hour. This means that forecasting flow for a node based only on data from previous time periods for nodes in the network may be insufficient to achieve a satisfactory result.

The multivariate nature of the network should be addressed in order to provide the best forecast performance. The relationships between the counting points may also be of interest, particularly when an unusual event occurs at one point that may be observed in downstream points. For example, a crash between counting points would likely result in decreased flow upstream and downstream of the incident, with possibly an above average flow once the accident has been cleared and the carriageway begins to clear.

1.3 Thesis plan

The structure of this thesis is as follows. The next chapter introduces the data set and performs some preliminary analysis before time series techniques are applied.

Chapter 2 summarises existing approaches to traffic forecasting with emphasis on traffic flow. Chapter 3 performs some preliminary analysis on the data available. Chapters 4 and 5 introduce the basics of the model adopted for this thesis. Chapter 4 describes the Dynamic Linear Model used for time series forecasting. Chapter 5 describes the Multiregression Dynamic Model and how it can be used to decompose a multivariate problem into conditionally independent univariate problems. Chapter 6 presents a mixture of standard results for the MDM and new techniques and analysis developed in this thesis for use with the DLM and MDM models.

In chapter 7 the MDM model is implemented for this application. The chapter also demonstrates the use of expert intervention in the model and compares model performance with and without such intervention. Chapter 8 introduces a modified MDM model and compares its performance to those in the previous chapter.

Chapters 9 and 10 apply different models to the same data set for comparison purposes. Chapter 9 applies independent DLM models to the problem and chapter 10 applies independent ARIMA models to the problem.

Finally, chapter 11 discusses the performance of the models applied in previous chapters and evaluates the MDM methodology as applied in this case. It also discusses possible areas for future research around MDMs in this type of application and in general.

Chapter 2 – Previous Approaches

Work looking at traffic forecasting falls into two main streams- mechanistic models that appear in transportation literature and models designed from a statistical basis. The former are concerned primarily with modelling traffic as a whole system in an explanatory manner. The latter typically concentrate on one aspect. However, there are relatively few approaches that use Bayesian techniques.

2.1 Previous non-Bayesian approaches

There have been many previous approaches to modelling traffic data (van Arem, Kirby et al. 1997).

However, in most cases the problem addressed is subtly different than here. For example, some work has centred on origin-destination matrices for vehicles passing through a traffic network (Camus, Cantarella et al. 1997) or the travel time through such a network (van Arem, van der Vlist et al. 1997). There is also work examining traffic state (for example, Wang and Papageorgiou 2005). These approaches do not address the problem at hand. However, there are relationships between traffic flow and the quantities considered in such work.

The most often used class of models are those based on Lighthill and Whitman (1955). The kinematic wave motion they describe gives a functional relationship between the flow q, the concentration k and the position x. The equation of continuity is of the first order:

$$\frac{\partial k}{\partial t}+\frac{\partial q}{\partial x}=0$$

which implies that the quantity in a (small) length changes at a rate equal to the difference between the inflow and the outflow. In the traffic application, this relates the flow, the vehicle density on the road and the road location. The relationship between flow and concentration for a fixed point x can be plotted as a flow-concentration curve. It is this curve that characterises the traffic flow at that point. The slope of the curve c is called the wave velocity.

$$c=\left(\frac{\partial q}{\partial k}\right)_{x\ constant}=c(k,x)$$

under the assumption that $q=q(k,x)$. If this c is less than the mean velocity at a point this means that velocity decreases with concentration, as observed in traffic networks.

This traffic conditions are modelled as a series of waves, defined by this flow-concentration curve. Under stationary conditions, this would be all that was required. Where the flow-concentration curve varies with the position x, the waves are no longer straight lines. However, the wave carrying a given flow q still has a predictable path. Under non-stationary conditions, discontinuities in this wave pattern occur, and will disperse either upstream or downstream according to the relative values of c and the mean velocity. The dispersion is akin to that observed with dynamic shock waves in gases, and as such discontinuities are referred to as shock waves.

Road junctions would be modelled as changes to the flow-concentration curve for a single road, and exceptional events as discontinuities. However, while the onset of an unusual event will produce a shock wave, it will also affect the flow-concentration curve of the road section.

In the application of this methodology to road traffic, q and k only have meaning as means, and as such can be measured directly. The flow-concentration curve has a maximum before congestion slows the vehicles sufficiently to reduce overall flow. The derivation of the curve is a based on observational data when the road is in a number of different traffic conditions.

This method only forecasts traffic flow in so much as it estimates the link between traffic flow, traffic density and the spatial positions of counting points. This explanatory approach, on its own, is of little use where the road conditions may fluctuate based on prevailing conditions. It does, however, reflect different road characteristics at different observation points.

This method requires extensive data collection at the counting points of interest before the model can be applied. The construction of the flow-concentration curves involves a certain amount of interpolation and would be opaque to someone not familiar with the methodology. Although expert opinion could be incorporated when unusual events occur, this would generally have to be on the level of the flow-concentration curves. Deriving the curves is complicated, but need only be done once for each counting point to apply for all possible time periods. Adverse conditions such as ice or rain would require different curves or some procedure for adjusting the existing curves. Finding all of this for even a modest road network could quickly become impractical. This methodology is essentially descriptive rather than predictive, concentrating on properties of the road segment such as maximum flow.

This model was also described in Richards (1956)- this method is generally known as the LWR model.

Yi et al (2003) apply a higher order extension to this technique, but such approaches have disadvantages (Daganzo, 1995, Heidemann 1999), but there are other models that use the LWR model as their basis. Some trace the effect of events upstream (see, for example, Zhang (1998), Hoogendoorn and Bovy (2000), Hoogendoorn and Bovy (2001) or Jiang, Wu and Zhu (2002)). These models look at the microscopic level and model how drivers behave with respect to the vehicles in front of them. There have been numerous other refinements to this approach, including Zhang (2001) and Zhang (2003). While studying the behaviour of traffic flow under incident conditions or under minor events in normal traffic conditions is informative, it might not necessarily translate directly into better forecasts for traffic flow in this situation where the time periods are relatively large. Models of this type typically deal with traffic merges by proposing some distribution scheme based on traffic demand (from upstream links) and supply (road characteristics), as seen in Jin and Zhang (2003). There are also criticisms specific to some deterministic car-following models in Nelson (1995). Many of these continuum models also require data or estimates beyond simple vehicle counts and the estimation can be difficult, as seen in Hurdle and Son (2000).

Supply and demand models and assignment models attempt to work out where the traffic will flow based on supply of road space and demand (which is from flow further upstream in network models). Akamatsu (1996) and Watling (1996) are two examples. In these models, there are multiple competing routes that drivers can select from in order to reach their destination.

Carey and Subrahmanian (2000) consider traffic flow on a congested stretch of road as a form of queue with a first-in-first-out property. This leads to a linear property in the model. Conservation of flow states that:

$$\sum_{j \in in(k)} \sum_{\tau=1}^{t} x_{j\tau t} + E_{kt} = \sum_{j \in out(k)} \sum_{\tau=t}^{T} x_{jt\tau} \; \forall k \in N, t=1,\ldots,T.$$

Where $x_{jt\tau}$ denotes the flow that enters link j at time t and leaves it at time τ, and E_{kt} denotes the exogenous demand (or in-flow) for node k at time t. The first-in-first out assumption places constraints on which $x_{jt\tau}$ are non-zero, and linear programming techniques can be used to solve the piecewise linear functions that Carey and Subrahmanian assume for the functions.

Neural net approaches have estimated traffic flow in addition to other quantities (Dougherty and Cobbett (1997) and Zhang et al (1997)). While this combined approach could be applied to this network, a neural net approach is a 'black box' system. Incorporating exogenous information is not practical unless the net can be trained to respond the correct way to such occasional data. As unusual events in this sort of situation are likely to be unique (in extent if not in form), it is unlikely that a neural net approach could adequately model this effectively. Additionally, a neural net is non-parametric and should the traffic network be substantively changed the net will need to be reformulated and trained again. However, neural networks offer the advantage that relationships between different aspects of the data (such as spatial relationships, transit time and current traffic conditions) are part of the neural net by default whereas statistical models must include them explicitly (Kirby, Watson et al. 1997).

There has been one approach (van der Zijpp and de Romph 1997) that provides an origin-destination approach coupled with a dynamic component that can report unusual events in the data. The origin-destination methodology was used as an interim component in order to forecast flow, traffic density and travel time. This dynamic approach has several advantages over the 'black box' approach in terms of interpretability, and has the advantage that the inner workings of the model have a real-world meaning outside the context of the model. The paper also attempts to cope with an unusual event by intervention in the model to account for it. However, the incorporation of this outside information is not a main feature of the model, and as such is an ad hoc response to the event. The solution of the model also involves an iterative algorithm which may not be suitable for a real-time system.

ARIMA methodology (Box and Jenkins 1976) has also been used extensively in traffic forecasting, for example in Eldor (1977), Gafarian et al (1977), Ahmed and Cook (1979) and Nihan and Holmesland (1980). However, there has been criticism of this approach (Okutani and Stephanedes 1984) suggesting that a simple moving average is just as effective. An ARIMA model consisting of a simple moving average with an additional seasonal component is used later in this thesis for comparison purposes.

Hjorth (2002) used a bootstrap technique on a road network with traffic lights. This stochastic method also considers route choice and travel time, rather than solely traffic flows. The presence of stop-go traffic at predetermined points also changes the behaviour of the network from a (usually) free-flowing motorway network.

Other macroscopic models have also been applied traffic forecasting- see (for example) Hilliges and Weidlich (1995) for an application to a traffic network and Sanwal et al (1996) for one to a freeway segment. Papageorgiou (1998) contends that there the descriptive accuracy of any macroscopic model in this application may never be as good as that of macroscopic models in other areas. However, Helbing et al show that some macroscopic models agree with microscopic models on traffic properties derived from them.

Characteristics of merging traffic in a motorway junction have been explored in Bunker and Troutbeck (2003)- where the proportion of delay between the main flow and the merging flow is estimated. However, this may be more useful for forecasting when traffic queues will occur or estimating journey times than for long time period flow forecasting.

The time gaps between vehicles in congested traffic flow (Banks 2003) can vary considerably between sites and indeed lanes, and do not seem to be correlated with common macroscopic factors such as vehicle speed. This implies that there are features of traffic flow at junctions that cannot adequately be found from such macroscopic factors. An attempt to forecast behaviour at such places might need to exhibit learning about these features.

Even where traffic flow is the primary concern, there is a dearth of Bayesian methods considered when comparing different methodologies. Smith and Demetsky (1997) consider ARIMA methodology, neural nets, non-parametric methods and historical average techniques but no Bayesian time-series.

There has also been work examining the behaviour of traffic networks solely under incident or congested conditions, for example Hounsell and Ishtiaq (1997), Cassidy and Maunch (2001) and Maunch and Cassidy (2002), which may inform any technique that is intended to operate under such conditions.

There are several aspects of structural models that might be of interest when considering this problem, even if they have not yet been applied specifically to this class of problem. For example, correlation in measurement errors between response and regressors has been considered in Schaalje and Butts (1993) and Reilman, Gunst and Lakshminarayanan (1985). Such a correlation may occur if flow in one area is regressed on flow in another. A multivariate structural normal model (as described in Fraser and Haq (1969) can be considered for such a network, without recourse to Bayesian methods.

2.2 The Bayesian approaches

Bayesian methods have been used to tackle this type of problem before. One method (Whittaker, Garside et al. 1997) used a state-space-based approach to forecast vehicle speed and flows. The model incorporates a term for exogenous factors which could be used to incorporate information regarding unusual events. The transition observation equations used were:

$$x_{t+1}=A_t x_t+Bu_t+q_t$$

$$y_t=Cx_t+r_t$$

where the u represent the exogenous factors. The decomposition of a complex network problem into a simpler series of problems is a key aspect both of this approach and that adopted in this thesis.

Bayesian methods were used by Yang and Davis (2002) to find classified mean traffic flow incorporating an AR(1) component- although in this case the

Bayesian component was to incorporate uncertainty in the classification of vehicles into the mean estimates.

Dynamic Linear Models or DLMs (West and Harrison 1999) fill many of the criteria required for this model. The Bayesian approach encapsulates the dynamic nature of the modelled system by updating current beliefs about the system as new data arrive. They are time series models which can be structured such that the parameters are readily interpretable and through the Bayesian methodology they can incorporate exogenous information when supplied in an appropriate form. This incorporation of outside data, known as 'intervention', can be achieved simply and without additional complication to the model. Formal methods for identifying periods of unusual activity also exist- which would aid the forecaster in a real-time system.

The DLM has been used to forecast traffic flows on a stretch of freeway (Tebaldi, West et al. 2002). A simple series of counting points, rather than a complicated network similar to that in this application, is modelled through a hierarchy of DLM models. Each counting point is regressed on the nearest upstream counting point at lags determined through analysis. However, the flow of traffic joining and leaving the freeway is modelled using a smoothed trend. For link i on day j at minute t, they used the equation:

$$y_{ijt} = s_{ij}(t) + b'_{ijt}\beta_{ij} + \epsilon_{ijt}$$

where $s_{ij}(t)$ was the smoothed entry and exit flow and b was a vector of (known) upstream flows. This can also be expressed in the usual notation for DLMs:

$$y_{ijt} = x'_{ijt}\theta_{ij} + \epsilon_{ijt}$$

The distributions for these quantities were assumed conjugate where appropriate, which gives:

$$\theta_{ij} \sim N\left(\theta_{ij} | \mu_i, \Sigma_i\right)$$

$$\mu_i \sim N\left(\mu_i | m_i, M_i\right)$$

$$\Sigma_i^{-1} \sim Wishart\left(\Sigma_i^{-1} | (\rho_i R_i)^{-1}, \rho_i\right)$$

$$\sigma_i^{-2} \sim \Gamma\left(\sigma_2^{-2} | \nu, \nu\tau^2/2\right)$$

Posterior distributions are calculated using Markov Chain Monte Carlo techniques which ran in under a minute. However, in larger networks the simulation may take markedly longer.

Many of these approaches (Bayesian or otherwise) are of limited use in this context as they use smaller time intervals for data collection. This has two qualitative effects on the model. Firstly, smaller time intervals give rise to noisier data as brief unusual activity is not smoothed away and patterns can be harder to discern as a result. Secondly, in sequenced regression models the lag between counting points may be non-zero, so forecasts can be made without reference to upstream nodes within the same time period. Where a vehicle can pass from one node to another within the same time period, as in the application considered in this thesis, the model must be able to regress on upstream nodes at lag 0.

Traffic flow shares several characteristics with fluid flow through pipes- one Bayesian technique for the latter was examined in Rougier and Goldstein (2001). However, in this case the Bayesian element was in the estimation of parameters that described the interaction between the pipeline and the fluid rather than forecasting future flow.

Multivariate DLMs may, at first glance, seem appropriate. Unfortunately, DLMs require certain variances in the system to be known, which will not be the case in general. Estimation of these quantities can be done easily in univariate models but is a non-trivial task in the multivariate case- which makes direct use of the DLM impractical.

Multiregression Dynamic Models or MDMs (Queen and Smith 1993) are a particular type of Bayesian multivariate dynamic model which can decompose a multivariate time series into a hierarchy of univariate time series, eliminating this problem. An MDM model can use a sequence of separate conditional regression DLMs to model its multivariate series. DLMs have the additional benefit that they can cope with missing data (which may be a feature of this sort of data) easily. An MDM approach was applied to this data before (Whitlock 1999). In this thesis, the application of the MDM to this problem is developed further.

Chapter 3 – Analysis of the Data

This chapter examines the raw data this thesis is using, preparatory to applying a suitable model to it. The relationships between the counting points and the time series collected at them will be examined and the form of the individual time series will be analysed.

3.1 Visual inspection

Consider the data for a few weeks at one of the counting points in figure 3.1.1. There are two obvious features of these data. Firstly, there is a pattern over the week, with the weekends showing less traffic than weekdays. Secondly, there is also a daily pattern, with morning and afternoon peaks. However, this pattern is not consistent over the entire week. In particular weekends seem to behave differently to weekdays. Similar patterns are observed at other counting points. This suggests that any model should have a seasonal pattern for days, with possibly different seasonal patterns for different days. Previous analysis of the data (Whitlock 1999) used principle components analysis to establish whether the daily patterns observed can be categorised. From that work three main categories emerge: Saturdays, Sundays/public holidays and weekdays. Most Mondays fall into the third category, although some behaved differently and were found to be public holidays. Fridays fall in the third category although they appear to have a slightly different daily pattern. These differences are all noted in the referenced analysis.

Although incorporating these different daily patterns into the model is not difficult, for clarity and simplicity this thesis uses only data on Tuesdays to

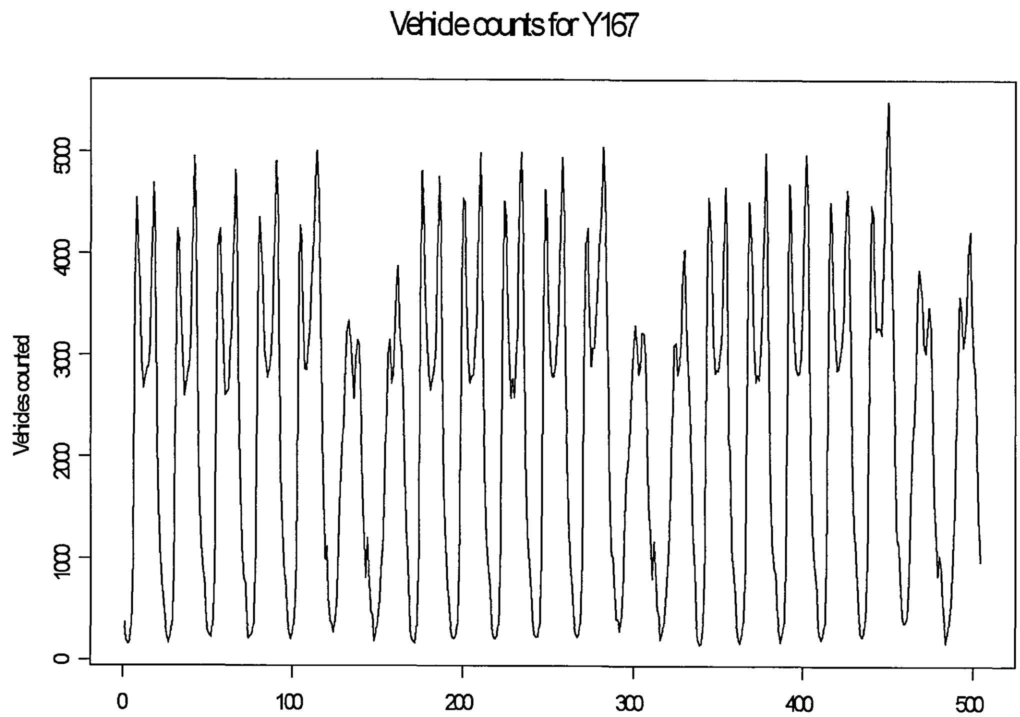

Figure 3.1.1 – Three weeks' data for counting point 167

Thursdays, to ensure all data fall within the third of those categories and thus exhibit a common daily pattern. This would not be adequate for a real-world implementation of the model but allows the model to be evaluated without the additional complication incurred by modelling the different day patterns. The multivariate aspect of the model is of most interest for this thesis, and so will be considered more thoroughly. The complication incurred by incorporating several daily patterns is not in the formulation of the model but in the easy analysis of the results.

3.2 Relationships between the sites

The geographical layout of the sites determines how traffic flows through the network. The distances between the counting points are such that a vehicle will normally travel all the way through the network in one time period. It is necessary to establish which counting points lead to others, and at which locations drivers make choices as to which route to follow. As motorway junctions include roundabouts, it is useful to make the simplifying assumption that drivers will follow the most direct route through the network and not, for example, drive around in circles or turn around and head back the way they came. Although some vehicles may behave this way, it is unlikely that such behaviour is common.

The first step is to plot what route is followed by vehicles through the network on the basis of where they enter the network and where they intend to leave the network. This can be determined from the geographical map (figure 1.1.1). For example, vehicles travelling south on the A282 that wish to leave the network via the southbound carriageway of the M25 pass three counting points- 167, 170a and 173a. Such a chain can be found for all entry and exit point pairs. There are routes through the network that do not pass any counting point- vehicles crossing, but not joining, the M25 at junction 1b either westbound or eastbound are not counted. Similarly, traffic crossing the M25 at junction 2 is also not counted. These vehicles are not an observable part of the traffic flow and do not enter the model. All other routes are shown in table 3.2.1.

Entry	Exit	Route
A282 southbound	M25 southbound	→167→170a→173a→
	Junction 1b	→167→168→
	A2 eastbound	→167→170b→171→
	A2 westbound	→167→170b→161→
Junction 1b	A282 northbound	→165→166→
	M25 southbound	→169→173b→
	A2 eastbound	→169→171→
	A2 westbound	→169→161→
A2 westbound	A282 northbound	→172→164b→166→
	M25 southbound	→172→173b→
	Junction 1b	→172→163→
A2 eastbound	A282 northbound	→162→164b→166→
	M25 southbound	→162→173b→
	Junction 1b	→162→163→
M25 northbound	A282 northbound	→164a→166→
	Junction 1b	→160→163→
	A2 eastbound	→160→171→
	A2 westbound	→160→161→

Table 3.2.1 – Origin-destination flows for the network

From table 3.2.1 it is possible to assemble a flow diagram showing how vehicles pass through the counting points in the network. This is shown in figure 3.2.1 where the counting points are represented by ovals and the flows between them by arrows. From the flow diagram it is possible to see all flows that lead to a particular counting point and all flows that lead away from it. All of the routes in table 3.2.1 can be traced through the diagram. Note that flows into and out of the network itself are listed in both the table and the diagram- it is important to track these quantities.

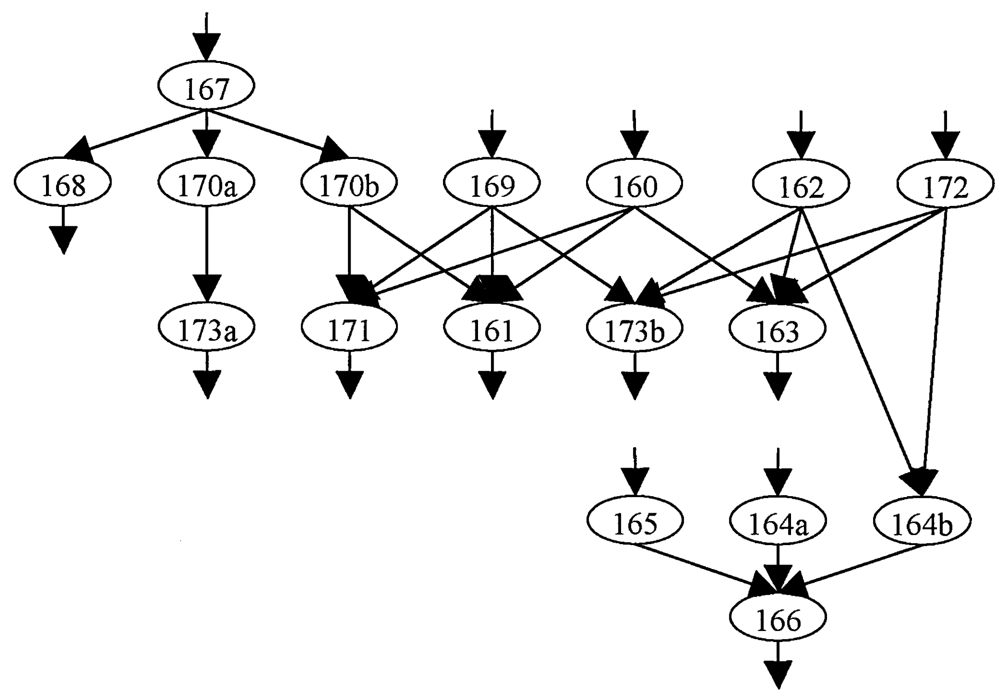

Figure 3.2.1 – Flow diagram showing how vehicles pass through the counting points in the network

An effective model will use these relationships to improve its forecast quality or exploit their relationships in some other way- such as by recognising that unusual patterns in one node are likely to be reflected in nodes it flows to and nodes that flow to it.

3.3 Missing data

Unfortunately there are missing data in this data set. Specifically, some counting points have no data collected for them. Without any data, it is difficult to attempt to incorporate these counting points in the model. They could be estimated using Markov Chain Monte Carlo techniques and have been in previous work (Whitlock and Queen 2000). However, this would not be practical in a real-time implementation. It is far simpler to remove such counting points from the flow diagram and use the reduced diagram as the basis for the model. There are no other missing data in this data set. Were there points where some of the data were missing, it might be possible to impute missing values another, such as using Bayesian networks (Di Zio, Scanu et al. 2004), or by using the DLM model's natural response to missing data.

3.4 Flow diagram incorporating missing data

The counting points for which no data are available are: 160, 166, 173a and 173b. Removing these sites affects the flow diagram. For example, the route →167→170a→173a→ becomes →167→170a→ when the missing site 173a is removed. Removing all the sites with missing data produces a new flow diagram, shown in figure 3.4.1.

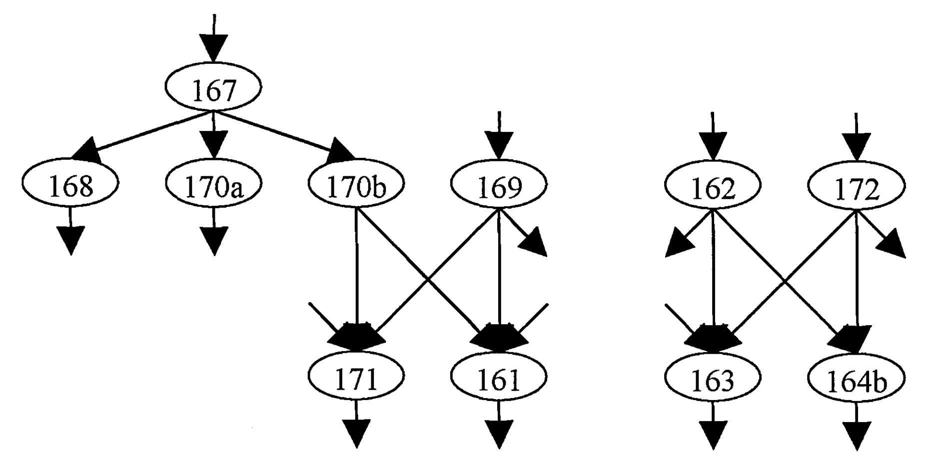

Figure 3.4.1 – Flow diagram for the network with missing and trivial sites removed

Notice that when these sites are removed from the flow diagram two other sites, 164a and 165, become disconnected from the rest of the network. They form no part of the hierarchical structure of the flow of traffic and are thus divorced from any multivariate structure. As this thesis aims to examine and model the multivariate nature of a traffic network they shall be dropped from the model. Additionally, the network is now subdivided into two separate subnetworks with no counting points in common.

The flow diagram is of great importance when building a model that incorporates the relationships between the counting points. The flow diagram will be revisited in chapter 6 when such a model is constructed.

Chapter 4 – Dynamic Linear Models

The model developed for the traffic network is based on Dynamic Linear Models (DLMs). This chapter introduces them and reproduces properties of them that will be used later.

4.1 Dynamic Linear Models

Dynamic Linear Models are a class of models commonly used as a Bayesian means of forecasting time series (West and Harrison 1999). A DLM is characterised by four matrices, indexed by time:

$$DLM\{F, G, V, W\}_t$$

This characterisation is used to construct two equations that define how the DLM process evolves:

Observation Equation: $$Y_t = F_t^T \theta_t + v_t \qquad v_t \sim N[0; V_t] \qquad \text{- 4.1.1}$$

System Equation: $$\theta_t = G_t \theta_{t-1} + w_t \qquad w_t \sim N[0; W_t] \qquad \text{- 4.1.2}$$

The two series of variables v_t and w_t are independent of all other aspects of the model, including each other and their own historical values. The (possibly multivariate) time series Y_t represents the quantity to be modelled, and the (again possibly multivariate) time series θ_t represents underlying quantities that the observations are influenced by. The G-matrix characterises how these underlying quantities evolve over time, while the F-matrix transforms these parameters into the observed quantities. The V- and W-matrices determine the random 'noise' associated with each equation. Noise in the observation equation is transient, whereas noise in the system equation affects the current level of the underlying parameter set.

By drawing inferences on the underlying parameters the forecaster can then forecast the observed quantities using the structure of the DLM. The system equation has a Markov property, so that only the current estimate of θ_t is required in order to forecast the series. The DLM model proceeds as follows:

Let the posterior distribution for the parameter θ_{t-1} be given by the equation:

$$(\theta_{t-1}|D_{t-1}) \sim N[m_{t-1}; C_{t-1}] \qquad \text{- 4.1.3}$$

where D_{t-1} denotes 'all information up to time t-1'. Note that often the conditioning on D_{t-1} is omitted from equations when dealing with the DLM; in general this condition is assumed unless specified otherwise. The structure of the model means that all this information can be encapsulated in this single distribution.

It is simple to derive the prior distribution for θ_t at time t-1 from the posterior for θ_{t-1} :

$(\theta_t|D_{t-1}) \sim N[a_t; R_t]$ where

$$\begin{aligned} a_t &= E(\theta_t|D_{t-1}) = E(G_t\theta_{t-1}+w_t) = E(G_t\theta_{t-1})+0 = G_tE(\theta_{t-1}) \\ a_t &= G_tm_{t-1} \end{aligned} \qquad \text{- 4.1.4}$$

$$\begin{aligned} R_t &= Var(\theta_t|D_{t-1}) = Var(G_t\theta_{t-1}+w_t) \\ R_t &= Var(G_t\theta_{t-1})+Var(w_t)+2cov(G_t\theta_{t-1},w_t) \\ R_t &= G_tVar(\theta_{t-1})G_t^T+W_t+0 = G_tC_{t-1}G_t^T+W_t \end{aligned} \qquad \text{- 4.1.5}$$

It is equally simple to derive the prior distribution of Y_t at time t-1 (called the one-step ahead forecast distribution) from the prior for θ_t :

$(Y_t|D_{t-1}) \sim N[f_t; Q_t]$ where

$$f_t = E(Y_t|D_{t-1}) = E(F_t^T\theta_t+v_t) = E(F_t^T\theta_t)+0 = F_t^TE(\theta_t) = F_t^Ta_t \qquad \text{- 4.1.6}$$

$$Q_t = Var(Y_t|D_{t-1}) = Var(F_t^T \theta_t + v_t)$$
$$Q_t = Var(F_t^T \theta_t) + Var(v_t) + 2cov(F_t^T \theta_t, v_t) = F_t^T Var(\theta_t) F_t + V_t + 0 \qquad \text{- 4.1.7}$$
$$Q_t = F_t^T R_t F_t + V_t$$

The above notation is standard in DLMs and will be used throughout this thesis.

Proof of the updating equations which calculate the posterior for θ_t after observing Y_t is not given here, but the result is reproduced below:

$$(\theta_t|D_{t-1}, Y_t) \sim N[m_t; C_t] \text{ where}$$

$$m_t = E(\theta_t|D_{t-1}, Y_t) = a_t + A_t e_t \qquad \text{- 4.1.8}$$

$$C_t = Var(\theta_t|D_{t-1}, Y_t) = R_t - A_t Q_t A_t^T \qquad \text{- 4.1.9}$$

and

$$e_t = Y_t - f_t \qquad A_t = R_t F_t Q_t^{-1}$$

Usually $\{D_{t-1} \quad Y_t\}$ is equivalent to D_t but in some circumstances D_t may contain additional information. One important example is when expert opinion is incorporated into the model.

There are two other quantities of interest with regard to a DLM which appeared in equations 4.1.8 and 4.1.9. The first is the one-step ahead forecast error e_t which indicates how well the model is performing. The second is known as the 'adaptive vector' A_t. The adaptive vector is of particular interest as not only is it used to find the posterior distribution of θ_t given Y_t, but it can be seen from equation 4.1.8 that it determines how quickly each element of θ_t will adapt to the observed value Y_t. The larger the value of A_t, the more m_t depends on e_t.

This gives the most general form of the DLM. In practice, there are several simplifications that can commonly be made that make the DLM even easier to implement, and several adjustments that are necessary for the DLM to be practical.

4.2 The univariate DLM

In a univariate DLM, the observation Y_t is a single value. This means that f_t, Q_t and V_t will also be single values and that F_t will be a vector with a length equal to the number of parameters in θ_t. The vector F_t and the matrix G_t are specified as part of the model and will often be constant over time. For the model to be workable it is essential that the series of posterior variances C_t for θ_t should converge. C_t is the variance of the estimate of θ_t and if it does not converge then the estimate is of little use. For constant F and G this has been proved for multivariate models that are 'observable' (Harrison 1997). A DLM model is observable if the observability matrix T given by

$$T = \begin{bmatrix} F^T \\ F^T G \\ \vdots \\ F^T G^{n-1} \end{bmatrix} \qquad \text{- 4.2.1}$$

is of full rank. Observability is a sufficient but not necessary condition for the convergence of C_t. The observability criterion depends only on the matrices chosen for F and G. This means that the forecaster can make a model observable, and thus convergent, by design.

4.3 Unknown system noise variance

The basic DLM requires the system noise variance, W_t, to be specified. There may not be a natural value to pick based on the problem, so an alternative method can be used. By choosing to specify W_t in terms of C_{t-1} (the posterior variance for

), the forecaster makes an assumption concerning the system noise variance. This is done using a discount factor δ, and defining W_t as follows:

$$W_t = G_t C_{t-1} G_t^T \frac{1-\delta}{\delta} \qquad 0 < \delta \leq 1 \qquad \text{- 4.3.1}$$

This now means that the prior variance for θ_t is given by:

$$R_t = G_t C_{t-1} G_t^T + G_t C_{t-1} G_t^T \frac{1-\delta}{\delta} = G_t C_{t-1} G_t^T \delta^{-1} \qquad \text{- 4.3.2}$$

This new derivation of R_t can be used throughout the rest of the prior equations and the updating equations. That particular form of W_t is chosen so R_t has the convenient form above. It is simple to implement and is meaningful in context as how the posterior variance for θ_{t-1} is 'discounted' to form the prior variance for θ_t. A value of δ =1 is equivalent to setting W_t=0, producing a static model for θ_t. High values of δ imply that there is little increase in the uncertainty and give low values of W_t. Low values of δ imply the opposite. Unlike specifying W_t explicitly, discounting is always relative to the current estimate of θ_t. Typical values for δ range from 0.8 to 1. A value smaller than this generally means the resulting forecast variances are too large for the model to be useful.

4.4 Unknown observation noise variance

Like W_t, the observation noise variance V_t needs to be specified as part of the DLM model but often this is not easily done. The DLM can be adapted to include 'variance learning'. It is assumed that V is constant but not known, and V is estimated through a parameter $\phi^{-1} = V$ that is updated in parallel with θ_t. The distribution of ϕ is Gamma, an appropriate Bayesian conjugate modelled through parameters S_t (the current point estimate of V) and n_t (a measure of the current precision of this point estimate). The updating equations for the variance learning portion of the model are as follows, following the same notation as previously:

Variance distribution: $$(\phi|D_{t-1}) \sim \Gamma\left[\frac{n_{t-1}}{2};\frac{n_{t-1}S_{t-1}}{2}\right] \qquad \text{- 4.4.1}$$

Updating equations: $$n_t = n_{t-1} + 1 \qquad \text{- 4.4.2}$$

$$S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right) \qquad \text{- 4.4.3}$$

Updated variance equation: $$(\phi|D_{t-1}, Y_t) \sim \Gamma\left[\frac{n_t}{2};\frac{n_t S_t}{2}\right] \qquad \text{- 4.4.4}$$

The distributions for the observation equation are Normal when conditioned on φ^{-1}, but the marginal distributions are Student-T distributions so with the same notation as before the prior distributions become:

Prior distribution: $$(\theta_t|D_{t-1}) \sim T_{n_{t-1}}[a_t; R_t] \qquad \text{- 4.4.5}$$

Forecast distribution: $$(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t; Q_t] \qquad \text{- 4.4.6}$$

The updating equation for C_t is changed as follows:

$$C_t = \frac{S_t}{S_{t-1}}\left(R_t - A_t Q_t A_t^T\right) \qquad \text{- 4.4.7}$$

As the model progresses n_t becomes large and the T distributions become increasingly similar to Normal distributions. However, as in the discounted system variance case, other aspects of the DLM are unaffected.

It is important to note that with Student-T distributions the matrices C_t and R_t are not variance matrices; rather they are the scale matrices for the T distributions. However, variances can be produced from these quantities, given n_t. This method of variance estimation works only with univariate models. Although methods exist for variance estimation in multivariate DLMs (West and Harrison 1999), they have restrictions on when they can be applied that makes them inappropriate for this traffic modelling problem. In particular they assume F_t, G_t and W_t are the same across all series which will not be the case in this application.

4.5 Final DLM

The DLM that will used in this model incorporates variance learning, discounting for W_t and constant G. Using the above work the DLM is as follows:

Observation equation: $$Y_t = F_t^T \theta_t + v_t \qquad v_t \sim N[0; V] \qquad \text{- 4.5.1}$$

System Equation: $$\theta_t = G\theta_{t-1} + w_t \qquad w_t \sim T_{n_{t-1}}[0; W_t] \qquad \text{- 4.5.2}$$

Information: $$(\theta_{t-1}|D_{t-1}) \sim T_{n_{t-1}}[m_{t-1}; C_{t-1}] \qquad \text{- 4.5.3}$$

$$(\phi_t|D_{t-1}) \sim \Gamma\left[\frac{n_{t-1}}{2}; \frac{n_{t-1}S_{t-1}}{2}\right] \qquad \text{- 4.5.4}$$

Discount assumption: $$W_t = GC_{t-1}G^T \frac{1-\delta}{\delta} \qquad 0 < \delta \leq 1 \qquad \text{- 4.5.5}$$

Forecasts: $$(\theta_t|D_{t-1}) \sim T_{n_{t-1}}[a_t; R_t] \qquad \text{- 4.5.6}$$

$$(Y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t; Q_t] \qquad \text{- 4.5.7}$$

where: $$a_t = G m_{t-1} \qquad R_t = G C_{t-1} G^T \delta^{-1}$$

$$f_t = F_t^T a_t \qquad Q_t = F_t^T R_t F_t + S_{t-1}$$

Updating equations: $$n_t = n_{t-1} + 1 \qquad \text{- 4.5.8}$$

$$S_t = S_{t-1} + \frac{S_{t-1}}{n_t}\left(\frac{e_t^2}{Q_t} - 1\right) \qquad \text{- 4.5.9}$$

$$m_t = a_t + A_t e_t \qquad \text{- 4.5.10}$$

$$C_t = \frac{S_t}{S_{t-1}}\left(R_t - A_t Q_t A_t^T\right) \qquad \text{- 4.5.11}$$

where: $e_t = Y_t - f_t$ $\qquad A_t = R_t F_t Q_t^{-1}$

The remainder of this chapter looks at two further aspects of DLMs that are important in this application- the introduction of seasonality and the incorporation of outside information.

4.6 Seasonality in DLMs

Where a time series exhibits apparent seasonality, a DLM can be specified such that this seasonality is encapsulated in the choice of F and G. This can be done independently of any variance discounting or estimation. There are three main methods of introducing seasonality into a DLM.

4.6.1 Seasonal factors

In the seasonal factors model, each time period within the seasonality has a single element of θ_t associated with it. Only one of these elements is used at a time, and they are cycled through time using the G matrix. These models are the most basic way of incorporating seasonality in the DLM model. F and G are of the following forms:

$$F = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad G = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & 0 & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

The dimensionality of the matrices is determined by the periodicity of the seasonality- in the case of hourly data F is 24×1 and G is 24×24.

4.6.2 Seasonal effects

The seasonal effects model is similar to that of the seasonal factors model, except that it assumes there is an underlying level, and the seasonality is modelled as deviations from that level. Note that the seasonal effects must sum to zero and additional constraints must be placed on the initial C_t matrix in order for this model to converge.

$$F=\begin{pmatrix}1\\1\\0\\\vdots\\0\end{pmatrix} \qquad G=\begin{bmatrix}1 & 0 & \cdots & \cdots & \cdots & 0\\ 0 & 0 & 1 & 0 & \cdots & 0\\ \vdots & \vdots & 0 & 1 & \ddots & \vdots\\ \vdots & \vdots & & \ddots & \ddots & 0\\ \vdots & 0 & & & 0 & 1\\ 0 & 1 & 0 & \cdots & \cdots & 0\end{bmatrix}$$

The dimensionality of the matrices is one greater than the periodicity of the seasonality- in the case of hourly data F is 25×1 and G is 25×25. The matrices are the same as for seasonal factors but with an extra initial row for F, and an extra initial row and initial column for the G matrix. This model is best used when the seasonality is believed to be a series of deviations from some underlying level, such as electricity use over a year. Seasonal effects provide estimates of both the underlying level and the deviations. Where it is believed there is no underlying level, or the level is of no particular interest, then the seasonal factor approach is generally preferred.

4.6.3 Fourier models

Seasonality can be expressed in terms of a Fourier decomposition. When the period of seasonality is odd, F and G have the form:

$$F=\begin{pmatrix} E_2 \\ E_2 \\ \vdots \\ E_2 \end{pmatrix} \qquad G=\begin{bmatrix} J_2(1,w) & 0 & \cdots & 0 \\ 0 & J_2(1,2w) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & J_2(1,hw) \end{bmatrix}$$

where for m=1,…,h:

$$E_2=\begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad J_2(1,mw)=\begin{bmatrix} cos(mw) & sin(mw) \\ -sin(mw) & cos(mw) \end{bmatrix}$$

When the period of the seasonality is even, F and G have the form:

$$F=\begin{pmatrix} E_2 \\ E_2 \\ \vdots \\ E_2 \\ 1 \end{pmatrix} \qquad G=\begin{bmatrix} J_2(1,w) & 0 & \cdots & 0 & 0 \\ 0 & J_2(1,2w) & \cdots & 0 & 0 \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & J_2(1,hw-w) & 0 \\ 0 & 0 & \cdots & 0 & -1 \end{bmatrix}$$

In both cases the value of h is dictated by the periodicity of the seasonality. A Fourier seasonal model of this sort is essentially a reparameterisation of a seasonal factors model. However, by eliminating harmonics that are deemed to have an insignificant impact on the model a smooth seasonal pattern can be achieved with fewer parameters when a Fourier model is used. However, the parameters will need to be transformed back again if the forecaster desires to examine them.

4.7 Expert intervention

In a time series model such as the DLM, a forecaster may receive information from outside the model that they wish to incorporate into their forecasts. The Bayesian basis makes this straight-forward in the DLM. The information can be included in the model in many ways which are given below. The forecaster may choose to intervene purely on the basis of prior information- incorporating information exogenous to the model. The forecaster may also intervene retroactively-choosing to apply a method of intervention just after the time point has been observed. The set of information that provokes the intervention is denoted by I_t. If there is some parameter associated with the intervention it can be denoted by i_t.

There are many ways that expert opinion can be included in the DLM, each well-suited to a particular situation. The forecaster must judge which is most appropriate for the circumstance.

In some circumstances unusual behaviour may follow a predictable pattern or have a consistent magnitude, in this case there can be formal methods for including such infrequent jumps as part of the model, after the change points in Makov (1983). However, these techniques would not be appropriate unless there was a reliable pattern in the jumps.

4.7.1 Disregarding data

If the expert believes that a particular observation is the result of factors not relevant to the modelling of the system, such as equipment error or a one-of-a-kind event, and they also believe that the data are not informative about future events, the time point can be treated as missing data. In this case:

$I_t = \{Y_t \text{ is to be treated as missing data}\}$

When missing data are encountered in a DLM, the prior estimates for θ_t become the posterior estimates without further change. The updating equations 4.5.8 to 4.5.11 become:

$$n_t = n_{t-1}$$

$$S_t = S_{t-1}$$

$$m_t = a_t$$

$$C_t = R_t$$

4.7.2 Transient change in level

The expert may believe that the observed value will be affected by some factor outside the model, but only for one time period. The expert can quantify by how much the observed value is expected to be changed:

I_t={Y_t is affected by a transient parameter i_t}

$$i_t \sim \left(f_t^*; Q_t^*\right)$$

The observation equation is changed for time t only to give:

$$Y_t = F_t^T \theta_t + i_t + v_t \qquad - 4.7.2.1$$

This leads to the forecasts:

$$\left(Y_t | D_{t-1}, I_t\right) \sim T_{n_{t-1}}\left[f_t; Q_t\right]$$

$$f_t = F_t^T a_t + f_t^* \qquad - 4.7.2.2$$

$$Q_t = F_t^T R_t F_t + Q_t^* + 2\,cov\left(F_t^T \theta_t, i_t\right) \qquad - 4.7.2.3$$

Note that consideration must be given to the covariance between i_t and $F_t^T \theta_t$, even if simply to assume it is zero. This intervention only affects a single time period as the θ parameter is unchanged by the intervention and the observation equation reverts to its original form in subsequent time periods.

4.7.3 Permanent change in level

If the expert believes that there will be a change in the underlying level at a certain time point, the system equation can be changed to reflect this change in level. This is done by inserting a parameter in the system equation in the same way as transient change was modelled by inserting a parameter into the observation equation. In this case:

I_t={Y_t is subject to a permanent change in level i_t}

$$i_t \sim \left[a_t^*, R_t^*\right] \qquad \text{- 4.7.3.1}$$

The system equation is changed for time t only to give:

$$\theta_t = G\,\theta_{t-1} + i_t + w_t \qquad \text{- 4.7.3.2}$$

This leads to the prior distribution for θ_t:

$$\left(\theta_t | D_{t-1}; I_t\right) \sim T_{n_{t-1}}\left[a_t; R_t\right]$$

$$a_t = G\,m_{t-1} + a_t^* \qquad \text{- 4.7.3.3}$$

$$R_t = G\,m_{t-1}\,G^T + R_t^* + 2\,cov\left(G\theta_{t-1}, i_t\right) \qquad \text{- 4.7.3.4}$$

As before, the system equation returns to its original form in subsequent time periods. Although the intervention only occurs in a single time point, the effects are long lasting as the level change has become part of the estimate of θ_t from that time forward.

4.7.4 Arbitrary intervention

The forecaster is at liberty to change any aspect of the model. The estimate of the observation noise variance, the discount factor δ, the estimate of ϕ or even the structural matrices F and G can be changed in response to outside information. The above methods merely describe useful ways of intervening that do not require major revision of the model and seamlessly integrate the intervention with the regular functioning of the model. Care must be taken when intervening that the intervention does not destroy useful information already contained in the model.

4.7.5 Formal monitoring

The question of whether to intervene does not depend solely on exogenous factors. If observed values lead the forecaster to believe that some unusual event has happened and the model is no longer adequately forecasting the system, they can choose to intervene to try to improve matters. This may provoke investigation of information outside the model to determine how to intervene, or intervention may be taken solely on the basis of the data. This monitoring of the data may be performed ad hoc by the forecaster, or a formal method may be employed to signal when the model is performing poorly. Formal monitoring methods similar to those used in quality control have been developed for DLMs. However, no formal methods have yet been developed for use with the models used in this thesis and thus formal monitoring is not considered here.

Chapter 5 – Multiregression Dynamic Models

A Multiregression Dynamic Model (Queen and Smith 1993) is a form of dynamic model that uses conditional independence and a causal driving mechanism within a system to decompose a multivariate time series into a network of univariate time series. Each univariate time series is conditional on some set of parent series. These univariate models can be updated independently and conditional forecasts for each univariate series can be found independently. Before the MDM can be introduced in detail, the Directed Acyclic Graph and conditional independence need to be introduced.

5.1 Directed acyclic graphs and conditional independence

A graph G(Y,E) is composed of a set of nodes Y and a set of edges E. A directed edge from Y_i to Y_j is denoted by $(Y_i, Y_j) \in E$. Graphs such as these are called directed graphs as the edges have a direction. If a directed graph is such that there are no 'loops' in the directed edges then it is a directed acyclic graph. Formally, a directed graph G is a directed acyclic graph (DAG) if there exists an ordering on Y such that:

$$\forall j, \forall i > j \qquad (Y_i, Y_j) \notin E$$

Figure 5.1.1 shows two diagrams, the first is a valid DAG, the second is not a valid DAG for two reasons- it has a loop (containing Y_1, Y_2 and Y_4) and it has an undirected edge (between Y_1 and Y_3).

An influence diagram is composed of a series of conditional independence statements and a particular graphical representation of those statements. For a set of

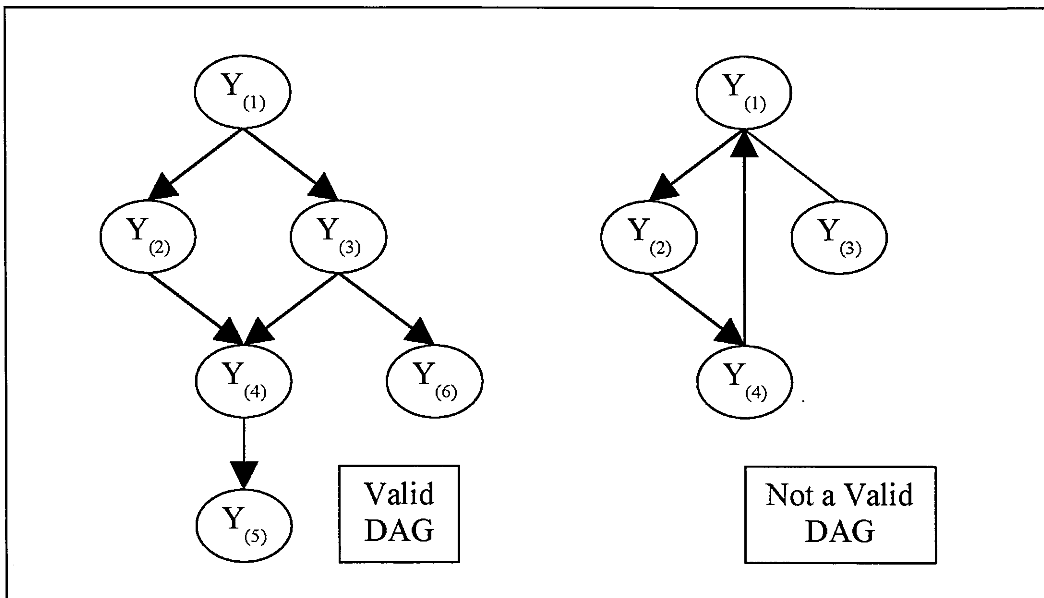

Figure 5.1.1 – Two example graphs, the first a valid DAG and the second not.

random variables $\{Y_1 \ldots Y_n\}$ the series of conditional independence statements is as follows:

$$Y_i \coprod \{Y_1 \ \ldots \ Y_{i-1}\} \backslash pa(Y_i) | pa(Y_i) \qquad i=2,\ldots,n \qquad , n \qquad -5.1.1$$

$\coprod$ denotes 'is independent of'

$\coprod$ denotes 'is independent of'

\ denotes set differencing

where $pa(Y_i)$ is called the 'parent set' of Y_i (Cowell, Dawid et al. 1999). Each element of this set is called a parent of Y_i. This statement says that Y_i is independent of all preceding variables except its parents, given the parent set. When defined across all Y_i it can be thought of as a function. These statements assert that Y_n is independent of all non-parent Y_i preceding it given those parent variables. This implies an ordering across the set of random variables. If a graph is constructed with edges from all the elements of $pa(Y_i)$ to Y_n then it can be seen that this forms a DAG. For example, if the parent sets of some set of nodes are as described here:

$$
\begin{aligned}
&pa(Y_2)=\{Y_1\}\\
&pa(Y_3)=\{Y_1\}\\
&pa(Y_4)=\{Y_2,Y_3\}\\
&pa(Y_5)=\{Y_4\}\\
&pa(Y_6)=\{Y_3\}
\end{aligned}
$$

then the following conditional independence structure is produced:

$$
\begin{aligned}
&Y_2 \neg\coprod Y_1\\
&Y_3 \coprod Y_2 | Y_1\\
&Y_4 \coprod Y_1 | Y_2, Y_3\\
&Y_5 \coprod Y_1, Y_2, Y_3 | Y_4\\
&Y_6 \coprod Y_1, Y_2, Y_4, Y_5 | Y_3
\end{aligned}
$$

$\neg$ denotes 'not'

This will generate the DAG shown in figure 5.1.1.

A certain degree of flexibility in the ordering of the nodes is generally present. The DAG and the conditional independence statements both specify the same structure, and together are considered to be the graph of an influence diagram. The benefits of formulating a DAG to describe these relationships are give in Pearl (1996).

5.2 Multiregression Dynamic Models

An MDM models a multivariate time series $Y_t = \{Y_t(1) \quad Y_t(2) \quad \dots \quad Y_t(n)\}$. Suppose that the ordering of the elements of this time series is such that the relationships between them can be represented heuristically by an influence diagram of the form shown in equation 4.1.1, i.e.:

$$Y_t(i) \coprod \{Y_t(1) \quad \dots \quad Y_t(i-1)\} \backslash pa(Y_t(i)) | pa(Y_t(i)) \qquad - 5.2.1$$

This influence diagram (and thus $pa(Y_t(i))$) remains the same for all t. It can be interpreted as the statement that the value of $Y_t(i)$ is independent of the values of all non-parent nodes at time t preceding it in the DAG given the values of its parents.

Suppose further that there is similar conditional independence through time, and that this structure can be represented heuristically as follows:

$$Y_t(i) \coprod \{Y^t(1) \quad \dots \quad Y^t(i-1)\} \backslash pa(Y^t(i)) | Y^{t-1}(i), pa(Y^t(i)) \qquad - 5.2.2$$

where $Y^k(r) = \{Y_1(r) \quad \dots \quad Y_k(r)\}$. The multivariate network can then be modelled as a series of univariate models using this conditional independence structure. Each node at time t has as its parent set any 'parent nodes' up to time t and its own history. Once the parent set is known, information about the other, non-parent nodes is of no additional use when forecasting $Y_t(i)$. This forms the basis for the MDM.

The MDM is a Bayesian forecasting system and, as such, the historical information $Y^{t-1}(i)$ can be encapsulated through the prior distribution of the set of parameters $\theta_t(i)$. Thus the conditional independence statements of equation 5.2.2 become:

$$Y_t(i) \coprod \{Y^t(1) \ \ldots \ Y^t(i-1)\} \setminus pa(Y_t(i)), Y^{t-1}(i) | pa(Y_t(i)), \theta_t(i) \qquad - 5.2.3$$

It is important to note that the MDM uses influence diagrams to represent conditional independence structure analogous to causality between the series. Variables that are hypothesised to be causally linked should be connected by a directed edge following the direction of causation (Wermuth and Lauritzen 1990). The direction of the edges in an MDM is important as it suggests a natural direction for the percolation of unusual events. If the MDM is to be used effectively when intervention is employed, this causative relationship.must be accurately portrayed. DAGs can represent the same conditional independence structure but with very different ordering. This is why the conditional independence statements (as thus, the ordering) are considered an important part of the conditional independence relationship.

For example, consider the two DAGs presented in figure 5.2.1. In both DAGs, $A \coprod C | B$. However, the causal relationships in the two DAGs are very different. The first implies a causal link from B to C, the second implies a causal link from B to A. One or the other, or both, may be true. The decision as to which DAG is the most appropriate should be made heuristically based on whichever best matches the inherent causality in the application in question. In the case of single direction traffic flow, as here, one direction would be more likely than the other.

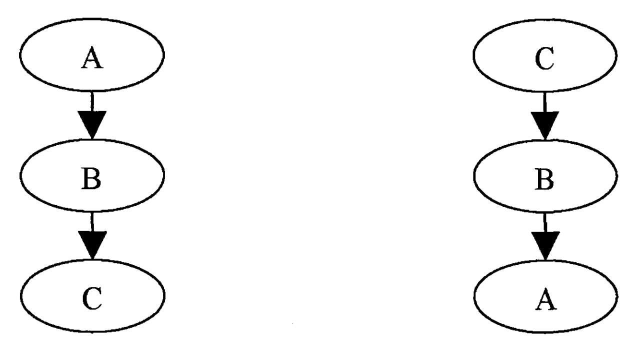

Figure 5.2.1 – Two superficially similar DAGs with different causative structure.

The MDM uses these influence diagrams to model each time series $Y_t(i)$ by a univariate model that regresses on its parents. It uses a system equation and an observation equation, as in the DLM, but they need not be linear. The observation equation in its most general form is:

$$Y_t(i)=f\left(F_t(i),\theta_t(i)\right)+v_t(i) \qquad \text{- 5.2.4}$$

where:

- The function $f(.)$ can be any known function.
- $v_t(i)$ need not be Normal
- $F_t(i)$ is a function of $pa(Y_t(i))$, but may also include exogenous variables.

When the function $f(.)$ is linear, this becomes the observation equation from a regression DLM. In this case, we have a linear MDM.

For each node, a system equation for an MDM can be specified as for any DLM with the following constraints.

- Across the whole model, all the $v_t(i)$ and $w_t(i)$ need to be independent of each other and other parts of the model through time.
- The $\theta_t(i)$ also need to be assumed independent of each other initially so that C_0 is block diagonal.
- G and W_0 must also be block diagonal.

The structure of the system equation means that the parameters $\theta_t(i)$ ... $\theta_t(n)$, if set independent of each other a priori, remain independent after the data are observed. Setting C_0 to be an appropriate block diagonal ensures independence a priori and thus independence for all time. This means that each variable $Y_t(i)$ can be modelled by an independent univariate model conditional on its parent set. Parameters for each univariate model can be updated separately, in sequence. The DAG structure defined across the time series has therefore been used to decompose the multivariate model into a series of separate conditional univariate models. Each of these models is a Bayesian dynamic model so the techniques available to them are directly applicable to the decomposed univariate models.

The observation equation for an element of a linear MDM is:

$$Y_t(i)=F_t(i)^T\theta_t(i)+v_t(i) \qquad v_t(i)\sim N\left[0;V_t(i)\right] \qquad -5.2.5$$

where, for example:

$$F_t(i)=\begin{pmatrix} pa\left(Y_t(i)\right)_{[1]} \\ pa\left(Y_t(i)\right)_{[2]} \\ \vdots \\ pa\left(Y_t(i)\right)_{[J]} \end{pmatrix}$$

and the notation $\bullet_{[j]}$ refers to the jth element of a vector or ordered set. $Y_t(i)$ is regressed on its parents from the DAG inside the same time period. This is similar to regressing $Y_t(i)$ on other observed values in previous time periods. However, regressing on values in the same time period requires more work to find marginal forecast distributions because the parents are not yet observed when the forecasts are made.

Seasonality can be incorporated into the model by constructing a $F_t(i)$ in the same way as for a seasonal DLM defined in section 4.6- except that the elements of $pa(Y_t(i))_{[j]}$ take the place of unit values. In the seasonal factors case this is:

$$F_t(i) = \begin{pmatrix} pa(Y_t(i))_{[1]} \\ 0 \\ \vdots \\ 0 \\ pa(Y_t(i))_{[2]} \\ 0 \\ \vdots \\ 0 \\ \vdots \\ pa(Y_t(i))_{[j]} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

A significant difference between a regression DLM and a non-regression DLM is that the parameter vector $\theta_t(i)$ is not a vector of absolute values itself but is scaled by the values of the regressors, in this case the parents of i. Discounting for W_t and variance learning for V can be used just as in the general DLM.

Some observed variables at the beginning of the network will have no parents. These can be labelled 'root nodes', as in figure 5.2.2. In the case of these root nodes any means of modelling a time series can be used. In fact there need not be a consistent approach throughout the network for root nodes, although it will generally be sensible to adopt one. If an MDM network uses a consistent method of modelling root nodes this could be indicated by calling it 'backed' by that modelling approach.

An MDM node may itself incorporate data exogenous to the MDM model (Queen and Smith 1993). The F vector for such a node would be:

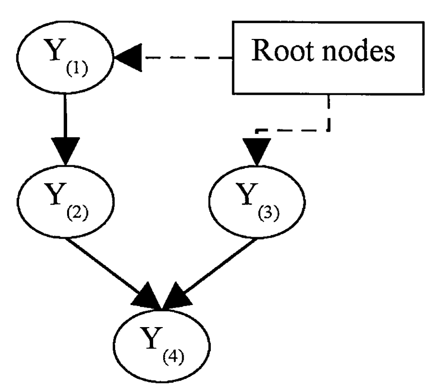

Figure 5.2.2 – An example DAG showing which nodes are root nodes.

$$F_t(i)=\begin{pmatrix} pa(Y_t(i))_{[1]} \\ \vdots \\ pa(Y_t(i))_{[j]} \\ x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix}$$

where the x_i represent exogenous variables. A constant term can be encapsulated by the '1'. The parameter vector would include subvectors of parameters for regressing on parent nodes, regressing on the exogenous variables and the constant regressor. An application may require an MDM node with this property- in the application here a node where traffic flows from one or more parents and from outside the network is an example. Because of this, 'backing' an MDM with DLMs is useful in that it would provide a consistent approach to flows into the network between root nodes and the kind of node described above.

5.3 Updating and forecasting in an MDM

The updating equations for the DLM are used unchanged in the linear MDM. The conditional forecast distributions for a linear MDM are Gaussian. However, these are conditioned on the parents for each node, which will not be observed when the node itself is being forecast. The marginal forecasts for the parents, unconditional on their own parents, are needed to obtain the marginal forecast for a node. Thus the marginal and joint forecast distributions for the MDM are highly non-Gaussian and generally intractable. This means acquiring marginal forecasts is more difficult than in DLMs. Fortunately, the updating equations are performed at the conditional level, meaning there is no additional difficulty in updating the model parameters. Although the forecast distributions are non-Gaussian, expected values, variances and covariances of marginal and joint forecasts can still be obtained, as will be shown in chapter 6.

To illustrate how this result holds, consider the DAG in figure 5.3.1. It is a DAG for a three-dimensional series with its parameters included. It is possible to verify that the parameter sets can be updated independently by constructing a moralised DAG. To construct a moralised DAG, all the parents for each node are joined with an undirected edge, and direction removed from the remaining edges. In a moralised DAG, two nodes are conditionally independent if all paths between them pass through the conditioning set. The moralisation of the DAG in figure 5.3.1 is shown figure 5.3.2.

Using the moralised DAG it is possible to verify that the parameters can be updated independently conditional on the observed values of the parent for that node. By tracing paths between nodes it can be seen that:

$$\theta_{(A)} \coprod \theta_{(B)}, \theta_{(C)} | A$$

$$\theta_{(B)} \coprod \theta_{(A)}, \theta_{(C)} | A, B$$

$$\theta_{(C)} \coprod \theta_{(A)}, \theta_{(B)} | B, C$$

This means that if the three parameters $\theta_{(A)}$, $\theta_{(B)}$ and $\theta_{(C)}$ are independent a priori, they remain so after observing A, B and C.

Although the flow of traffic suggests a natural direction of causation in this application, in some circumstances the direction of causation may not be unequivocal. In these conditions chain graph models (Lauritzen and Richardson 2002) are an appropriate representation of the causal structure but without

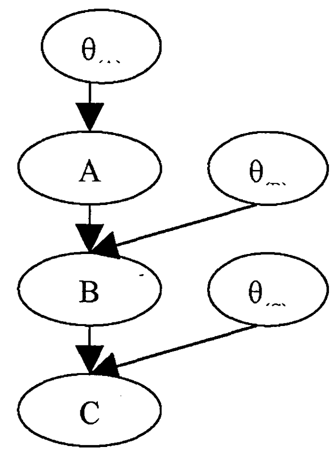

Figure 5.3.1 – Example DAG for an MDM network

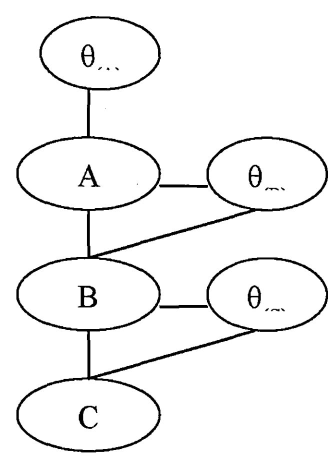

Figure 5.3.2 – Example moralised DAG for an MDM network

modification can't be used as the basis for an MDM as the non-directed link does not indicate which variable should be regressed on which. In circumstances where the application does not suggest a natural direction of causation, exploratory techniques exist to determine which model is most appropriate (for example Consonni and Leucari 2001).

5.4 Intervention in the MDM

A linear MDM is a series of regression DLMs, and hence all techniques for intervention in a DLM can be employed in a linear MDM. The parameters are regression parameters as opposed to absolute levels but mechanically there is no change needed. Additionally, the conditional independence structure means that the MDM model is an expert system, and techniques appropriate to expert systems can be considered here (Spiegelhalter, Dawid et al. 1993). A real-time forecasting system for this application will require intervention to be performed relatively quickly, incorporating a variety of sources of information. Should these sources need to be combined to provide a suitable composite before intervention is applied to the MDM, the method in Faria and Smith (1997) is appropriate. The standard techniques of intervention for the DLM can incorporate such information or pooled opinion satisfactorily.

5.4.1 Intervention by tiers

The hierarchical structure does introduce some additional effects on intervention that offer significant advantages over a multivariate DLM of the same network. When intervention is performed at a node in an MDM network, then this intervention affects the forecast for that node. The forecast of that node is then used as the basis for the forecasts for its children. The intervention percolates through the entire network down from where it was performed for no additional effort. This is in stark contrast to a multivariate DLM where intervention must be performed across all affected quantities. As intervention generally requires non-trivial decisions concerning manner and extent, the gain is marked. Figure 5.4.1.1 shows an example DAG for a series and illustrates where this gain is made.

Assume that the modeller determines (through whatever means) that intervention is required at node A. If the causal structure in the DAG is valid, then the nodes descended from A (B, C and E) share this period of unusual activity and will likely require intervention also. However, when intervention is performed on node A, the forecast including the intervention is used as the basis of the forecast for B, and similarly for nodes B and C then C and E. This means that a single intervention at node A may be sufficient to correct unusual activity in all four nodes. However, there may be unusual activity at the same time in nodes B, C and E that is not related to that at node A- so it is necessary to check whether intervention is needed after the intervention at A has been performed. However, there is no need to check nodes lower in the network before checking node A. In a collection of univariate models or a multivariate model, the assessment of whether intervention is necessary needs to be performed simultaneously across nodes. In this example, unusual activity at node A would require intervention at nodes A, B, C and E, a fourfold increase in the number of interventions (and hence the number of decisions concerning extent of intervention) that have to be made.

The approach deals with each layer completely before moving on to the next- and hence will be called here 'intervention by tiers'. In intervention by tiers the modeller begins by examining the behaviour of the nodes at the very highest level of

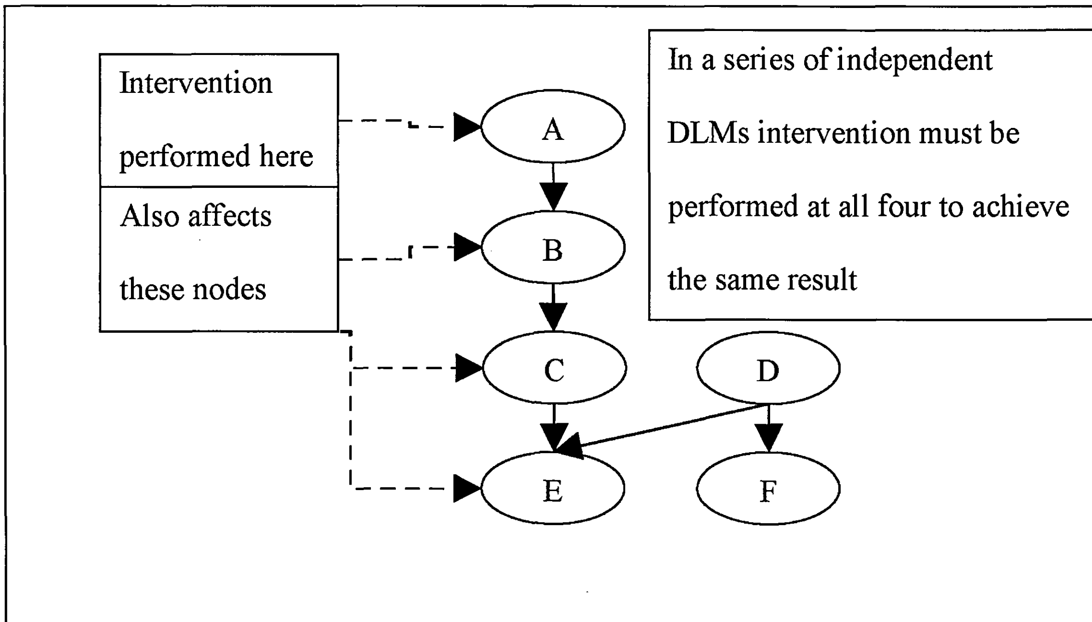

Figure 5.4.1.1 – Example DAG showing how intervention filters through tiers in the model

the network, which will consist only of root nodes that are at the very 'top' of the model. Once the modeller is satisfied that intervention is not necessary or has applied intervention then the next level down through the hierarchy is examined. This level will consist of any children of nodes in the first level and any root nodes that will be required before the following level can be examined. This continues until all nodes have been examined in this way and intervention applied where necessary. In figure 5.4.1.1, the first phase involves performing intervention at A, if any. The second phase examines B after intervention at A and sees if any intervention needs to be performed for it. The third phase examines C and D, in the light of any previous intervention, and the final phase would see if intervention was still needed for E or F after all previous intervention.

Every node in the network is examined to see if intervention is required (as would be expected) but the number of interventions that have to made may be lower by virtue of the MDM network structure. Even in the worst case intervention by tiers only requires as much work as intervening simultaneously in a multivariate DLM model.

An important ramification of intervention by tiers is that the network hierarchy should reflect any causal drive through the system, as discussed in section 5.2. Should this not be the case, intervention at one node will percolate to a node that is unaffected by the event itself, requiring further intervention to correct the erroneous change.

Selecting a DAG for an MDM that accurately reflects the causative links between nodes will avoid problems such as erroneous intervention. The causation is what enables intervention by tiers to work and should be preserved if the MDM is to be effective.

In some circumstances, an event in the network may cause counter-causal effects. For example, a road accident will not only affect nodes further down the network but will also cause traffic queues that may impinge on nodes further up the network. However, despite the counter-causal nature of the event it can be dealt with as though the cause of the event was where the end of the queue was. The forecaster can intervene at the furthest upstream point of the queue, and if necessary intervene at the point of the accident to reflect possibly different behaviour at the two locations. The nodes downstream of the accident site inherit the intervention as normal, and the nodes between the end of the queue and the accident site inherit the intervention of the queue end. Where a node has more than one parent, the upstream extension of the event will require different intervention at each point, and possibly intervention when they join. Even in the worst case, however, it will only involve as much intervention as a model that doesn't exploit the causative structure. The forecaster may have historical information that suggests a pattern in such cases, which may reduce the work required to intervene. If intervening proactively, the forecaster will have to assess where the queue is likely to extend to and whether additional intervention is required at the accident site in addition to the end of the queue. When intervening retroactively, both of these can be established from the forecast errors in the data.

In the data used here, there is no contextual information as to what causes any period of unusual activity and thus whether any of the events are counter-causal.

To illustrate how this works in practice, consider the plots of forecast errors for a number of nodes in this data set shown in figure 5.4.1.2

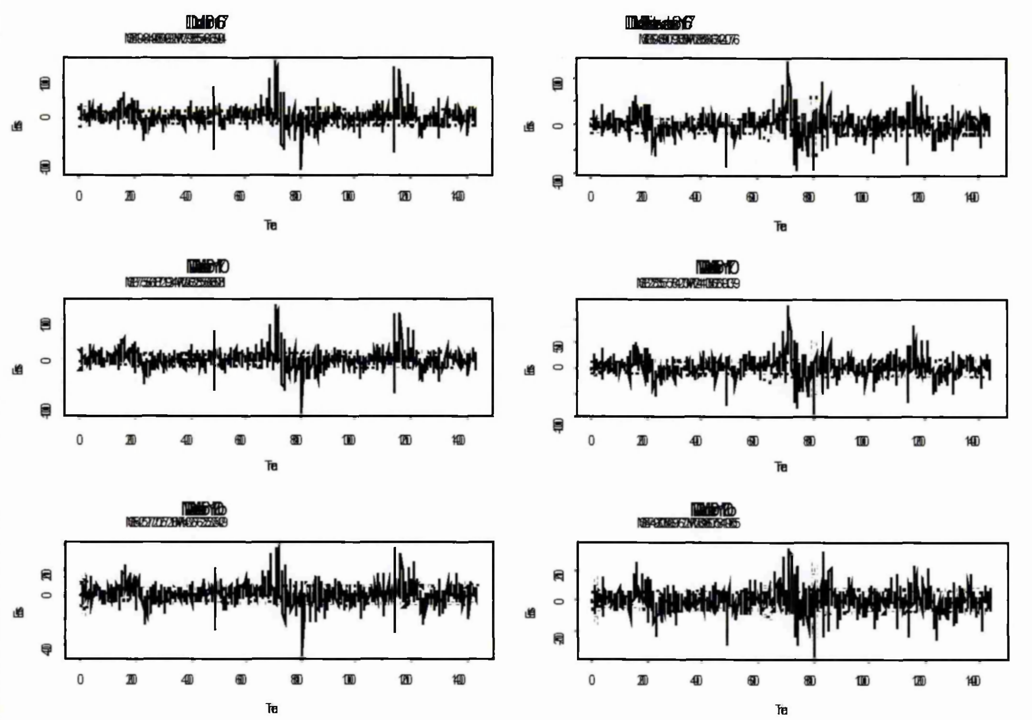

Figure 5.4.1.2 – Forecast error plot showing how intervention filters through the MDM

Those on the left show the pattern of forecast errors without intervention. The two lower nodes are downstream of the top node, and the unusual activity in this period is common to all three. The forecaster, however, only considers the top node and decides to intervene. The plots on the right are after intervention has been performed at the top node only. The plot for the top node after intervention shows the gain made there. Having done so, the forecaster then considers the error plot for the second node after the intervention in the first (middle right plot) and decides that intervention is not required at this point. Finally, they consider the final forecast error plot and again decide that no further intervention is necessary for this event. More details on the interventions performed here are given in chapter 7.

Chapter 6 – Old and New Results in DLMs and MDMs

This chapter presents results and ideas for DLMs and MDMs beyond the introduction to them in previous chapters. Some of the results are standard and others are presented here for the first time. They are commingled because some of them are heavily interrelated. There are three main areas of these results. The first is the concept of a 'deterministic twin'- building on deterministic nodes used in MDMs previously. The second deals with forecasts in an MDM network including a deterministic twin. The third section concerns a new technique of intervention in DLMs and linear MDMs.

6.1 Deterministic twins

This section considers a particular kind of deterministic node in an MDM network. The need for it arose through applying the linear MDM to the traffic network considered here but its use is far more general.

If the value of a parent node must be equal to the sum of its children (because of the nature of the application, say) then this places a restriction on the model parameters. It implies that the value of a child node is deterministic given those of the other children and the parent. In the case where a node has only two children, it is only necessary to model one child as an MDM node. The other can be simply calculated and forecast from its 'twin' and its parent. This deterministic node, while part of the MDM network, is neither a regular MDM node nor a root node. It is a deterministic node which here will be called a *deterministic twin*. The forecasts for the parent and the child can be used to generate forecasts for this deterministic twin. The twin itself can be indicated in a DAG with the usual notation of a square box for deterministic nodes. To illustrate, consider figure 6.1.1 where A must equal the sum of B and C. In this case, the parameters $\theta_{(B)}$ and $\theta_{(C)}$ must sum to 1. However, if nodes B and C were modelled separately with parameters $\theta_{(B)}$ and $\theta_{(C)}$, then there would be no restriction on their sum being 1. Attempting to model both separately with this restriction would result in a DAG where the parameters were not independent, breaking the requirement of the MDM. By setting one of them to be a deterministic twin, the restriction is in place in a way that does not break the assumptions of the MDM. The revised DAG with a deterministic twin is shown in figure 6.1.2.

These deterministic twins are of interest because a network may have deterministic twin nodes with children. Obtaining forecasts for deterministic twins permits modelling their children. It is also of note that the choice of which child is modelled directly and which is left as a deterministic twin is not forced by the model. Even where the deterministic twin is not observed it is forecast as a consequence of forecasting its parent and sibling.

In the case where a parent node has more than two children that sum to it, the conditional independence of the parameters is not preserved even if a deterministic twin is used. Whichever children are modelled directly do not have their parameters modelled independently. In this case, it is necessary to break the parent/child structure into a number of levels, each of which has a parent with only two children. Consider the candidate DAG in figure 6.1.3 where A is the sum of B, C and D. The parameters $\theta_{(B)}$ and $\theta_{(C)}$ will not be independent, because their sum is constrained. Instead, the structure must be broken into a series of tiers as shown in figure 6.1.4.

As in the two-child case, the choice of which child is modelled directly and where it is modelled is not forced by the model. The nodes created for this approach

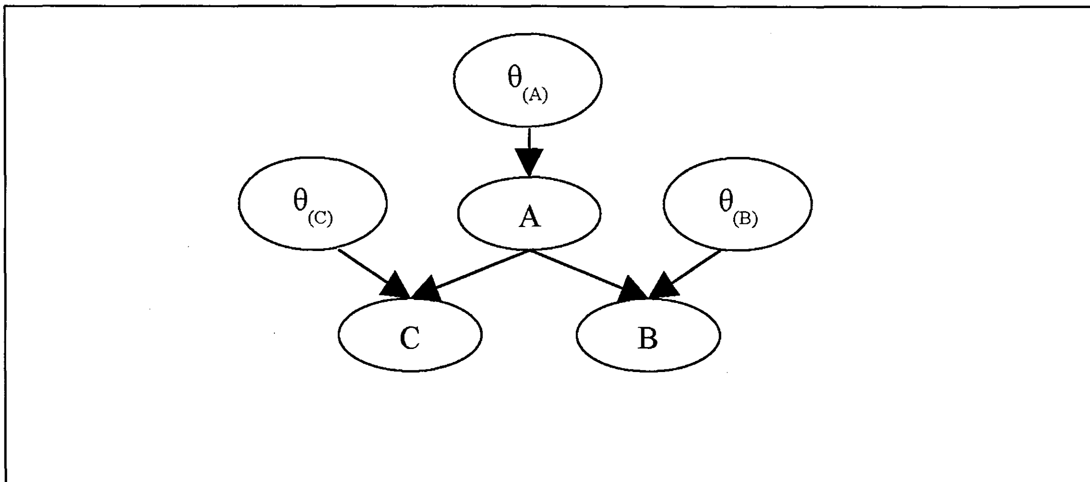

Figure 6.1.1 – Incorrect DAG for an MDM network with constrained children.

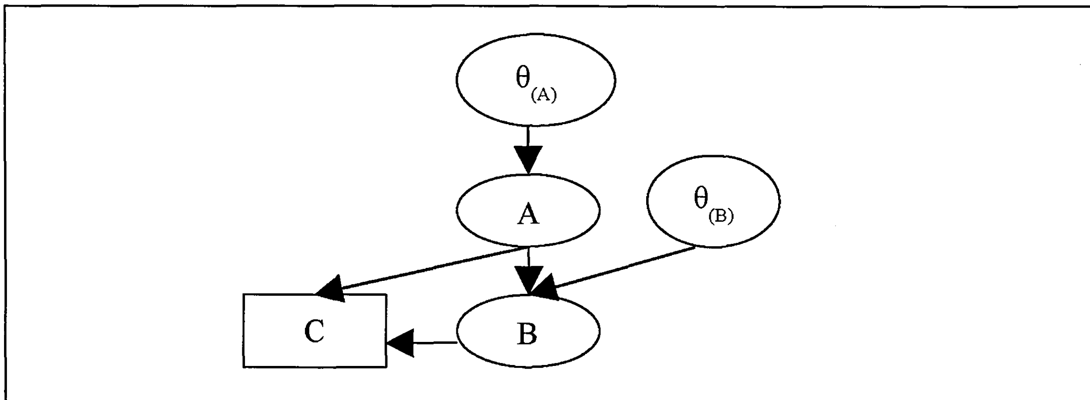

Figure 6.1.2 – DAG for an MDM network with a deterministic twin C

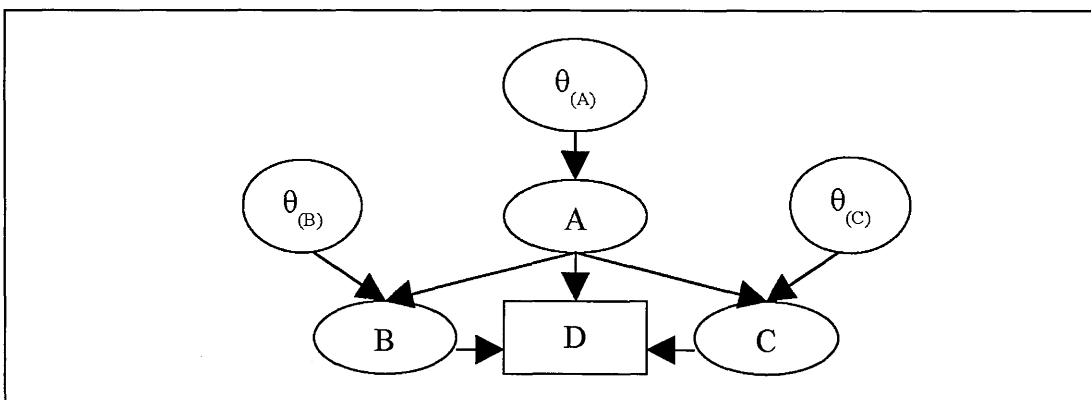

Figure 6.1.3 – Example of a parent with three children summing to it- but without correct independent updating.

need to be observed for the approach to work, but it is always possible to select quantities that will be observed as long as the children themselves are observed.

Nodes with a greater number of children simply require more tiers of the order of O(log n). Whenever such a restriction exists, a parent node may only ever have two children, one of which must be a deterministic twin.

Where a restriction exists that is not equality, then similarly the parent/child structure should be broken down. For example, in a flow network two children must not sum to greater than their parent. In this case, the difference between the sum of the children and the parent is not directly observed and may be of no interest to the forecaster but is still calculated from the other nodes and forms part of the MDM structure.

Note that this approach deals with restrictions on the observed quantities, not those imposed directly on the parameters. Additionally, this approach is not of great utility where there are multiple parents, as it is not clear what the quantity for the deterministic twin should be. This decomposition is similar to the competitive market models of Queen (1997), however in this case the decomposition must be strictly binary. It is also superficially similar to the assignment problem considered in Cargnoni et al. (1997) where the approach to categorical data used conditioning on the known and total constant.

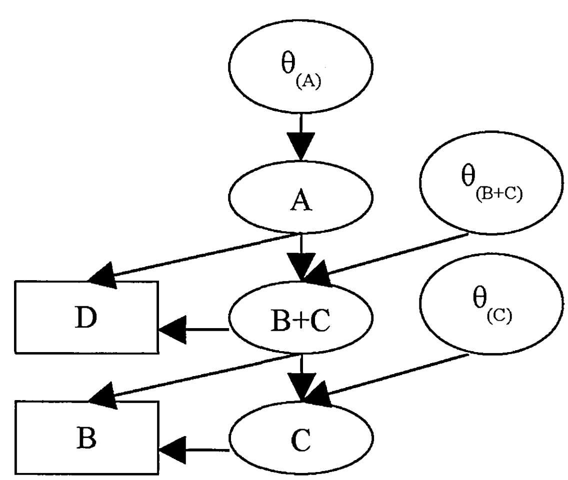

Figure 6.1.4 – Example DAG showing how three siblings are broken into tiers

6.2 One-step ahead forecasts

Using standard DLM methods it is possible to find one-step ahead forecasts for an MDM node conditional on its parent nodes. However, it is more useful to be able to calculate the marginal forecasts. Derivation of these marginal forecasts is straightforward (Queen and Smith 1993). The results are derived below in the notation established in chapter 4.

For node i, denote the marginal forecasts by $f_t(i)^* = E\left[Y_t(i) | D_{t-1}\right]$ and $Q_t(i)^* = V\left[Y_t(i) | D_{t-1}\right]$.

Root nodes are not MDM nodes and thus $f_t(i)^* = f_t(i)$ and $Q_t(i)^* = Q_t(i)$ for each such node i.

For two random variables X and Y there is the general result:

$$
\begin{aligned}
&E[X] = E[E(X|Y)] \\
&\Rightarrow f_t(i)^* = E[f_t(i)] \\
&f_t = E\left[F_t(i)^T a_t(i)\right] \\
&f_t = E\left[F_t(i)^T\right] a_t(i)
\end{aligned}
$$

Define $F_t(i)^{*T}$ to be $\left[f_t\left(pa(i)_{[1]}\right)^* \quad 0 \quad \cdots \quad 0 \quad \cdots \quad f_t\left(pa(i)_{[n]}\right)^* \quad 0 \quad \cdots \quad 0 \quad \cdots\right]$

In other words, $F_t(i)^*$ is of the same form as $F_t(i)$, but with the marginal forecast means for node i's parents in place of the observed values- $F_t(i)^* = E\left[F_t(i) | D_{t-1}\right]$.

Then:

$$f_t(i)^* = F_t(i)^{*T} a_t(i) \qquad \text{- 6.2.1}$$

Now consider $Q_t(i)^*$. For any two random variables:

$$Var[X]=E[Var(X|Y)]+Var[E(X|Y)]$$
$$\Rightarrow Q_t(i)^*=E[Q_t(i)]+Var[f_t(i)]$$

For the first half of the right hand side:

$$E[Q_t(i)]=E\left[F_t(i)^T R_t(i) F_t(i)+S_{t-1}(i)\right]$$
$$E[Q_t(i)]=E\left[F_t(i)^T R_t(i) F_t(i)\right]+S_{t-1}(i)$$

As $R_t(i)$ is a covariance matrix it is positive definite and can be decomposed as follows:

$$R_t(i)=B_t(i) B_t(i)^T$$

where $B_t(i)$ is non-singular. This gives:

$$E[Q_t(i)]=E\left[F_t(i)^T B_t(i) B_t(i)^T F_t(i)\right]+S_{t-1}(i)$$

Using the identity for two vectors a and b:

$$a^T b=trace\left(a b^T\right)$$

and setting $a^T=F_t(i)^T B_t(i)$ and $b=B_t(i)^T F_t(i)$ this becomes:

$$E\left[trace\left(B_t(i)^T F_t(i) F_t(i)^T B_t(i)\right)\right]+S_{t-1}(i)$$
$$=trace\left(E\left[B_t(i)^T F_t(i) F_t(i)^T B_t(i)\right]\right)+S_{t-1}(i)$$

Further, using the identity for a vector x that:

$$Var[x]=E\left[x x^T\right]-E[x]E[x]^T$$
$$\Rightarrow E\left[x x^T\right]=Var[x]+E[x]E[x]^T$$

this becomes:

$$trace\left(Var\left[B_t(i)^T F_t(i)\right]+E\left[B_t(i)^T F_t(i)\right]E\left[B_t(i)^T F_t(i)\right]^T\right)+S_{t-1}(i)$$
$$=trace\left(B_t(i)^T Var\left[F_t(i)\right]B_t(i)+B_t(i)^T E\left[F_t(i)\right]E\left[F_t(i)\right]^T B_t(i)\right)+S_{t-1}(i)$$
$$=trace\left(B_t(i)^T Var\left[F_t(i)\right]B_t(i)+B_t(i)^T F_t(i)^* F_t(i)^{*T} B_t(i)\right)+S_{t-1}(i)$$

The matrix $Var(F_t(i))$ is easy to construct from the marginal variances of the parent nodes and the marginal covariances between them. The former are the $Q_t(i)^*$ for the parents and the latter is considered in the next section.

For the second half of the equation for $Q_t(i)^*$:

$$Var[f_t(i)]=Var[F_t(i)^T a_t(i)]$$
$$Var[f_t(i)]=a_t(i)^T Var[F_t(i)]a_t(i)$$

and the variance of $F_t(i)$ has already been described above.

Finally, the marginal variance is:

$$Q_t(i)^*=trace\left[B_t(i)^T\left(Var(F_t(i))+F_t(i)^*F_t(i)^{*T}\right)B_t(i)\right]+a_t(i)^T Var(F_t(i))a_t(i)+S_{t-1}(i)$$

-6.2.2

Notice that where a node has more than one parent, finding $Var[F_t(i)]$ requires the marginal covariances between the parent nodes. One-step ahead forecasts for deterministic twin nodes can be found from the marginal forecast moments of their associated nodes. This requires finding the covariance between a node and its parent. Finding the covariance between two nodes in an MDM network is covered in the next section.

The same forecast quantities can be found for a deterministic twin as follows: Let $Y_t(j)$ be the deterministic twin of $Y_t(i)$, with parent $Y_t(p)$.

$$f_t(j)^*=E[Y_t(j)]=E[Y_t(p)-Y_t(i)]=f_t(p)^*-f_t(i)^* \qquad -6.2.3$$

$$Q_t(j)^*=Var[Y_t(j)]=Var[Y_t(p)-Y_t(i)]$$
$$Q_t(j)^*=Q_t(p)^*+Q_t(i)^*-2cov(Y_t(p),Y_t(i)) \qquad -6.2.4$$

Again, a covariance is needed in order to calculate the forecast variance.

6.3 One-step ahead covariance matrix

From the above, it can be seen that in MDM networks where nodes have more than one parent, or where the forecasts for deterministic nodes are of interest, it is necessary to calculate the covariance between nodes. It is possible to calculate the covariances iteratively. There is a general form (Queen and Smith 1993) that finds the covariances between all the nodes. However, in some circumstances the covariances of individual components of a node with other parts of the model may be useful in themselves. Even if covariances are only of concern in calculation of the one-step ahead variances, it is necessary to find them. The iterative method provided in Queen and Smith can be simplified to the point where it becomes a single line equation for each node as follows:

For an MDM node $Y_t(i)$ with parents $j_1, \ldots, j_k$, begin by partitioning $F_t(i)$ and $\theta_t(i)$ as follows:

$$\begin{aligned} F_t(i)^T &= \left[F_t(i, j_1)^T \quad F_t(i, j_2)^T \quad \cdots \quad F_t(i, j_k)^T\right] \\ \theta_t(i)^T &= \left[\theta_t(i, j_1)^T \quad \theta_t(i, j_2)^T \quad \cdots \quad \theta_t(i, j_k)^T\right] \end{aligned} \qquad \text{- 6.3.1}$$

and further defining the analogous quantities for the prior mean of $\theta_t(i, j_n)$:

$$a_t(i)^T = \left[a_t(i, j_1)^T \quad a_t(i, j_2)^T \quad \cdots \quad a_t(i, j_k)^T\right]$$

In the case of a node that also has an inflow from outside the model, recall that this is functionally identical to having a constant parent. This means that each component of decomposition 5.3.1 refers uniquely to one parent of (or flow in to) the node. Define:

$$Y_t(i, j) = F_t(i, j)^T \theta_t(i, j)$$

The observation equation can now be rewritten as:

$$Y_t(i)=Y_t(i,j_1)+Y_t(i,j_2)+\ldots+Y_t(i,j_k)+v_t(i) \qquad \text{- 6.3.2}$$

To find the variance between this node $Y_t(i)$ and any other node A, apply the fact:

$$cov(A,Y_t(i))=\sum_{n=1}^{k} cov(A,Y_t(i,j_n)) \qquad \text{- 6.3.3}$$

Equations 5.3.2 and 5.3.3 are in accordance with the superposition principle for DLMs (West and Harrison 1999). So, to find the covariance between $Y_t(i)$ and any other node in the network, it is only necessary to find all the covariances between the k components $Y_t(i,j_n)$ and this other node. This also simplifies the finding of covariances in general as it is only necessary to consider nodes with single parents.

To illustrate, consider the example DAG given in figure 6.3.1. In this DAG let $Y_t(7)=\begin{pmatrix}Y_t(5) & Y_t(6)\end{pmatrix}\theta_t(7)+v_t(7)$.

Also let $\theta_t(7)=\begin{pmatrix}\theta_t(7,5)\\ \theta_t(7,6)\end{pmatrix}$.

Then $Y_t(7,5)=Y_t(5)\theta_t(7,5)$ and $Y_t(7,6)=Y_t(6)\theta_t(7,6)$, so the covariance between $Y_t(1)$ and $Y_t(7)$, for example, can be written as:

$$cov(Y_t(7),Y_t(1))=cov(Y_t(7,5),Y_t(1))+cov(Y_t(7,6),Y_t(1)) \qquad \text{- 6.3.4}$$

Now it is possible to calculate the covariances between any nodes using this decomposition technique. Firstly, ensure that the ordering of the nodes obeys the following:

$$\forall i,j<n \qquad i<j\Rightarrow j\notin pa(i)$$

If this is not the case, it is simple to relabel the nodes so that it is true. There is always such an ordering as it is a requirement in order for a DAG to be drawn to

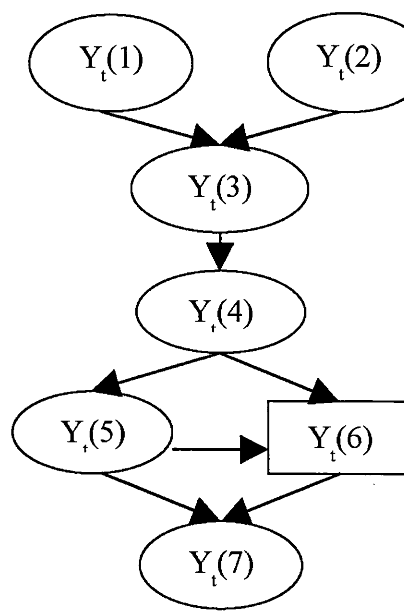

Figure 6.3.1 – Example DAG for use in covariance examples

express the conditional independence statements in chapter 5. The general iterative equation given in Queen and Smith (1993) is:

$$\begin{bmatrix} cov(Y_t(1), Y_t(r)) \\ cov(Y_t(2), Y_t(r)) \\ \vdots \\ cov(Y_t(r-1), Y_t(r)) \end{bmatrix} = E\begin{bmatrix} Y_t(1)Y_t(r) \\ Y_t(2)Y_t(r) \\ \vdots \\ Y_t(r-1)Y_t(r) \end{bmatrix} - \begin{bmatrix} E[Y_t(1)]E[Y_t(r)] \\ E[Y_t(2)]E[Y_t(r)] \\ \vdots \\ E[Y_t(r-1)]E[Y_t(r)] \end{bmatrix}$$

$$= E\left[\begin{pmatrix} Y_t(1) \\ Y_t(2) \\ \vdots \\ Y_t(r-1) \end{pmatrix} \cdot E[Y_t(r) | Y_t(1) \quad \cdots \quad Y_t(r-1)]\right] - \begin{bmatrix} E[Y_t(1)]E[Y_t(r)] \\ E[Y_t(2)]E[Y_t(r)] \\ \vdots \\ E[Y_t(r-1)]E[Y_t(r)] \end{bmatrix}$$

- 6.3.5

but this can be simplified as shown in the following theorem and corollaries.

Theorem 6.3.1:

In a linear MDM, let $cov(Y_t(i), Y_t(j))$ be known for some i and j. Let $Y_t(j)$ be the sole parent of $Y_t(r)$. Then:

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), F_t(r))a_t(r)$$

$F_t(r)$ will be in terms of $Y_t(j)$, and as $cov(Y_t(i), Y_t(j))$ is known $cov(Y_t(i), Y_t(r))$ can be found.

Proof:

To find $cov(Y_t(i), Y_t(r))$, consider row i of equation 6.3.5:

$$cov\left(Y_t(i),Y_t(r)\right)=E\left[Y_t(i)\cdot E\left[Y_t(r)\|Y_t(1)\ \ \dots\ \ Y_t(r-1)\right]\right]-E\left[Y_t(i)\right]E\left[Y_t(r)\right]$$

Now,

$$E\left[Y_t(r)\|Y_t(1)\ \ \dots\ \ Y_t(r-1)\right]=F_t(r)^T a_t(r)$$

and

$$E\left[Y_t(r)\right]=E\left[E\left[Y_t(r)\|Y_t(1)\ \ \dots\ \ Y_t(r-1)\right]\right]$$

so:

$$\begin{aligned}
&cov\left(Y_t(i),Y_t(r)\right)\\
&=E\left[Y_t(i)\cdot F_t(r)^T a_t(r)\right]-E\left[Y_t(i)\right]E\left[E\left[Y_t(r)\|Y_t(1)\ \ \dots\ \ Y_t(r-1)\right]\right]\\
&=E\left[Y_t(i)\cdot F_t(r)^T a_t(r)\right]-E\left[Y_t(i)\right]E\left[F_t(r)^T a_t(r)\right]\\
&=E\left[Y_t(i)\cdot F_t(r)^T\right]a_t(r)-E\left[Y_t(i)\right]E\left[F_t(r)^T\right]a_t(r)\\
&=cov\left(Y_t(i),F_t(r)^T\right)a_t(r)
\end{aligned}$$

as required. □

Corollary 6.3.2:

Continuing from Theorem 6.3.1, so that $cov\left(Y_t(i),Y_t(j)\right)$ is known and $Y_t(j)$ is the sole parent of $Y_t(r)$, assume the MDM structure is aseasonal or uses seasonal factors. Then:

$$cov\left(Y_t(i),Y_t(r)\right)=cov\left(Y_t(i),Y_t(j)\right)a_t(r)_{[1]}$$

where, as usual, $a_t(r)_{[1]}$ denotes the first element of the vector $a_t(r)$. Thus, all that is needed to find $cov\left(Y_t(i),Y_t(r)\right)$ is the covariance between $Y_t(i)$ and $Y_t(j)$ (which is known) and the prior point estimate of the parameter set $\theta_t(r)$.

Proof:

$F_t(r)$ is defined in terms of $Y_t(j)$, as $Y_t(j)$ is the sole parent of $Y_t(r)$. Where seasonal factors are used or the model is aseasonal the first element of $F_t(r)$ is simply $Y_t(j)$. Any other elements of $F_t(r)$ (in the seasonal factors case) are constant 0, hence the covariance between them and $Y_t(i)$ is also 0. Theorem 6.3.1 then simplifies to the above.

□

Corollary 6.3.3:

Continuing from Corollary 6.3.2, consider the special case where i = j, so that $Y_t(i)$ is the sole parent of $Y_t(r)$:

$$cov(Y_t(i), Y_t(r)) = Q_t(i)^* a_t(r)_{[1]}$$

which is a very simple means of calculating the covariance between a parent $Y_t(i)$ and a child $Y_t(r)$.

Proof:

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j)) a_t(r)_{[1]}$$ by Corollary 6.3.2.

$$= Var[Y_t(i)] a_t(r)_{[1]}$$ since i=j

$$= Q_t(i)^* a_t(r)_{[1]}$$

as required.

□

Corollary 6.3.4

Let $Y_t(r)$ be a deterministic twin of $Y_t(s)$, with $Y_t(j)$ as their single parent.

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j)) - cov(Y_t(i), Y_t(s))$$

This means that the covariance between deterministic twins and any other node can be found from the covariances between their parent and sibling and the node required.

Proof:

$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j) - Y_t(s))$ from the definition of a deterministic twin

$= cov(Y_t(i), Y_t(j)) - cov(Y_t(i), Y_t(s))$

as required.

□

Corollary 6.3.5

Consider the special case of Corollary 6.3.4 where i=j (the covariance between a deterministic twin and its parent). Then:

$$cov(Y_t(i), Y_t(r)) = Q_t(i)^* - cov(Y_t(i), Y_t(s))$$

Proof:

Follows from applying i=j to Corollary 6.3.4.

□

Corollary 6.3.6

Consider the special case of Corollary 6.3.4 where i=s (the covariance between a deterministic twin and its sibling). Then:

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j)) - Q_t(i)^*$$

Proof:

Follows directly from Corollary 6.3.4 given i=s.

□

Corollary 6.3.7

Consider Corollary 6.3.4, where $Y_t(r)$ is the deterministic twin of $Y_t(s)$ with $Y_t(j)$ as their sole parent. If the MDM node uses seasonal factors or is aseasonal, then:

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j))(1 - a_t(s)_{[1]})$$

which is analogous to Corollary 6.3.2.

Proof:

$$cov(Y_t(i), Y_t(r)) = cov(Y_t(i), Y_t(j)) - cov(Y_t(i), Y_t(s)) \quad \text{from Corollary 6.3.4.}$$

$$= cov(Y_t(i), Y_t(j)) - cov(Y_t(i), Y_t(j))\, a_t(s)_{[1]} \quad \text{from Corollary 6.3.2.}$$

$$= cov(Y_t(i), Y_t(j))(1 - a_t(s)_{[1]})$$

as required.

□

Corollary 6.3.8

Given a node with k parents $j_1, \ldots, j_k$:

$$Y_t(r) = Y_t(r, j_1) + Y_t(r, j_2) + \cdots + Y_t(r, j_k) + v_t(r)$$

decomposed as in equation 6.3.2, and assuming that $cov(Y_t(i), Y_t(j_n))$ is known for all n, then:

$$cov\left(Y_t(i), Y_t(r)\right) = \sum_{n=1}^{k} cov\left(Y_t(i), F_t(r, j_n)\right) a_t(r, j_n)$$

Further more, if $Y_t(r)$ is aseasonal or uses seasonal factors:

$$cov\left(Y_t(i), Y_t(r)\right) = \sum_{n=1}^{k} cov\left(Y_t(i), Y_t(j_n)\right) a_t(r, j_n)_{[1]}$$

Proof:

Use equation 6.3.3 to show that:

$$cov\left(Y_t(i), Y_t(r)\right) = \sum_{n=1}^{k} cov\left(Y_t(i), Y_t(r, j_n)\right)$$

Applying Theorem 6.3.1 to each member of the summation gives:

$$cov\left(Y_t(i), Y_t(r)\right) = \sum_{n=1}^{k} cov\left(Y_t(i), F_t(r, j_n)\right) a_t(r, j_n)$$

as required. The second result stems from applying Corollary 6.3.2 to the first result.

So the covariance $cov\left(Y_t(i), Y_t(r)\right)$ can be found simply from the covariances of the parents of $Y_t(r)$- $cov\left(Y_t(i), Y_t(j_n)\right)$ for $n = 1, \ldots, k$ - as long as they are known.

□

These corollaries allow the calculation of covariances between all nodes in the network given any covariances between entry points very simply. It is thus possible to populate a covariance matrix for all $Y_t(i)$ in this manner, assuming that all covariances between entry points (root nodes and the entry components of MDM nodes with an in-flow) are known.

6.3.1 Example

To illustrate, an example is given of the equations in practice. Consider once again the DAG shown in figure 6.3.1. If it is known that $cov\left(Y_t(1), Y_t(2)\right)=0$ then it is possible to calculate the covariance matrix. For simplicity of illustration assume no seasonality and a number of parameters equal to the number of parents for each MDM node. The observation equations for the network are presented below:

$$
\begin{aligned}
&Y_t(1)=\theta_t(1)+v_t(1)\\
&Y_t(2)=\theta_t(2)+v_t(2)\\
&Y_t(3)=\begin{pmatrix}Y_t(1) & Y_t(2)\end{pmatrix}\begin{pmatrix}\theta_t(3,1)\\ \theta_t(3,2)\end{pmatrix}+v_t(3)\\
&Y_t(4)=Y_t(3)\,\theta_t(4)+v_t(4)\\
&Y_t(5)=Y_t(4)\,\theta_t(5)+v_t(5)\\
&Y_t(6)=Y_t(4)-Y_t(5)\\
&Y_t(7)=\begin{pmatrix}Y_t(5) & Y_t(6)\end{pmatrix}\begin{pmatrix}\theta_t(7,5)\\ \theta_t(7,6)\end{pmatrix}+v_t(7)
\end{aligned}
$$

To illustrate how to calculate a covariance, consider $cov\left(Y_t(7), Y_t(1)\right)$. Recall equation 6.3.4:

$$cov\left(Y_t(7), Y_t(1)\right)=cov\left(Y_t(7,6), Y_t(1)\right)+cov\left(Y_t(7,5), Y_t(1)\right)$$

For the first part of the right hand side:

$cov\left(Y_t(7,6), Y_t(1)\right)=cov\left(Y_t(6), Y_t(1)\right)a_t(7,6)$ from Corollary 6.3.8.

$cov\left(Y_t(6), Y_t(1)\right)=cov\left(Y_t(4), Y_t(1)\right)-cov\left(Y_t(5), Y_t(1)\right)$ from Corollary 6.3.4.

$cov\left(Y_t(5), Y_t(1)\right)=cov\left(Y_t(4), Y_t(1)\right)a_t(5)$ from Corollary 6.3.2.

therefore:

$$cov\left(Y_t(7,6), Y_t(1)\right)=cov\left(Y_t(4), Y_t(1)\right)\left(1-a_t(5)\right)a_t(7,6)$$

For the second part of the right hand side:

$$cov\left(Y_t(7,5),Y_t(1)\right)=cov\left(Y_t(5),Y_t(1)\right)a_t(7,5)$$ from Corollary 6.3.8.

therefore:

$$cov\left(Y_t(7,5),Y_t(1)\right)=cov\left(Y_t(4),Y_t(1)\right)a_t(5)a_t(7,5)$$

So:

$$\begin{aligned}&cov\left(Y_t(7),Y_t(1)\right)\\&=cov\left(Y_t(4),Y_t(1)\right)\left(1-a_t(5)\right)a_t(7,6)+cov\left(Y_t(4),Y_t(1)\right)a_t(5)a_t(7,5)\\&=cov\left(Y_t(4),Y_t(1)\right)\left(\left(1-a_t(5)\right)a_t(7,6)+a_t(5)a_t(7,5)\right)\end{aligned}$$

Now:

$$cov\left(Y_t(4),Y_t(1)\right)=cov\left(Y_t(3),Y_t(1)\right)a_t(4)$$ from Corollary 6.3.2.

$$cov\left(Y_t(3),Y_t(1)\right)=cov\left(Y_t(3,1),Y_t(1)\right)+cov\left(Y_t(3,2),Y_t(1)\right)$$

from Corollary 6.3.8.

$$cov\left(Y_t(3,1),Y_t(1)\right)=Q_t(1)^* a_t(3,1)$$ from Corollaries 6.3.3 and 6.3.8.

and:

$$cov\left(Y_t(3,2),Y_t(1)\right)=cov\left(Y_t(2),Y_t(1)\right)a_t(3,2)$$ from Corollary 6.3.8.

$$cov\left(Y_t(3,2),Y_t(1)\right)=0$$ since $cov\left(Y_t(1),Y_t(2)\right)=0$

therefore:

$$cov\left(Y_t(4),Y_t(1)\right)=Q_t(1)^* a_t(3,1)a_t(4)$$

So finally:

$$cov\left(Y_t(7),Y_t(1)\right)=Q_t(1)^* a_t(3,1)a_t(4)\left(\left(1-a_t(5)\right)a_t(7,6)+a_t(5)a_t(7,5)\right)$$

which is a simple equation in terms of prior means for regression parameters and the forecast variance for $Y_t(1)$.

All the covariances produced by this method are simple and easy to calculate. Where the values are found recursively, the production of the equations becomes easier. If the full covariance matrix is not required but specific covariances are needed to calculate marginal variances, this technique is swifter than a fully recursive technique.

6.4 The expanded covariance matrix

It can be seen that the approach for calculating covariances described in section 6.3 makes extensive use of the decomposed covariances, for example $cov(Y_t(7,6), Y_t(3))$. Further, the same decomposition technique can be applied in order to find $cov(Y_t(7,6), Y_t(3,2))$, for example. When calculating the covariance matrix it is possible to store these values as they may contain information useful to the forecaster, for example when predicting the effect of structural changes to the model. These covariances between decomposed parts of the model can be inserted into another matrix introduced here called the *expanded covariance matrix*.

In this matrix each row and column represent the covariance between one part of one node and one part of another (possibly the same) node. Each node has one row and one column associated with it for each term in equation 6.3.2. Each node is decomposed by the number of regressors in F_t plus the observation noise term. It can be populated using the method above and can contain more precise information about the flow of traffic than the standard covariance matrix.

Adopt the following notation:

Define Σ_t to be the one-step ahead covariance matrix for all the $Y_t(i)$.

Define $\overline{\Sigma_t}$ to be the one-step ahead expanded covariance matrix for all the $Y_t(i)$.

Define $\overline{F_t}$ to be such that:

$$\overline{F_t^T}\begin{vmatrix} Y_t(1,1) \\ \vdots \\ Y_t(1,k_1) \\ v_t(1) \\ \vdots \\ Y_t(n,1) \\ \vdots \\ Y_t(n,k_n) \\ v_t(n) \end{vmatrix} = \begin{pmatrix} Y_t(1) \\ \vdots \\ Y_t(n) \end{pmatrix}$$

In other words, $\overline{F_t}$ is defined such that it maps from the decomposed components of all $Y_t(i)$ to the $Y_t(i)$ themselves.

It is then possible to find Σ_t from $\overline{\Sigma_t}$ and $\overline{F_t}$.

$$\Sigma_t = \overline{F_t}\,\overline{\Sigma_t}\,\overline{F_t^T} \qquad \text{- 6.4.1}$$

This means that if the expanded covariance matrix is found, then the covariance matrix can be calculated from it easily.

Consider the first three nodes in the example DAG in figure 6.3.1. The expanded covariance matrix has seven rows and columns. Assume the two root nodes have a simple constant $F_t(i)=1$ (so that $Y_t(i)=\theta_t(i)+v_t(i)$) and that $cov(Y_t(1), Y_t(2))=0$. The seven quantities that form the rows and column of the matrix are then $\theta_t(1)$, $v_t(1)$, $\theta_t(2)$, $v_t(2)$, $Y_t(3,1)$, $Y_t(3,2)$, $v_t(3)$.

As an example of calculating the entries for the expanded covariance matrix, consider $cov(\theta_t(1), Y_t(3,1))$. Following the decomposition in equation 6.3.2, $\theta_t(i)$ is $Y_t(1,\bullet)$, where here • represents a parent that has the constant value of 1.

$cov(\theta_t(1), Y_t(3,1)) = cov(Y_t(1,\cdot), Y_t(3,1))$ as shown above.

$cov(\theta_t(1), Y_t(3,1)) = cov(Y_t(1,\cdot), Y_t(1))a_t(3,1)$ from Corollary 6.3.8.

$cov(\theta_t(1), Y_t(3,1)) = cov(Y_t(1,\cdot), Y_t(1,\cdot)+v_t(1))a_t(3,1)$ from equation 6.3.2.

$cov(\theta_t(1), Y_t(3,1)) = cov(Y_t(1,\cdot), Y_t(1,\cdot))a_t(3,1) + cov(Y_t(1,\cdot), v_t(1))a_t(3,1)$

$cov(\theta_t(1), Y_t(3,1)) = Var(Y_t(1,\cdot))a_t(3,1)$ as the $v_t(i)$ are independent of $Y_t(1,\bullet)$.

$cov(\theta_t(1), Y_t(3,1)) = Q_t(1,\cdot)^* a_t(3,1)$

However, $Q_t(1,\bullet)^*$ in this case is $Var(\theta_t(1))$, so:

$cov(\theta_t(1), Y_t(3,1)) = R_t(1)a_t(3,1)$

Now consider $cov(v_t(1), Y_t(3,1))$. By the same approach:

$cov(v_t(1), Y_t(3,1)) = cov(v_t(1), Y_t(1))a_t(3,1)$ from Corollary 6.3.8.

$cov(v_t(1), Y_t(3,1)) = cov(v_t(1), Y_t(1,\cdot)+v_t(1))a_t(3,1)$ from equation 6.3.2.

$cov(v_t(1), Y_t(3,1)) = cov(v_t(1), Y_t(1,\cdot))a_t(3,1) + cov(v_t(1), v_t(1))a_t(3,1)$

$cov(v_t(1), Y_t(3,1)) = Var(v_t(1))a_t(3,1)$ as the $v_t(i)$ are independent of $Y_t(1,\bullet)$.

$cov(v_t(1), Y_t(3,1)) = V_t(1)a_t(3,1)$

The entire matrix can be populated in this manner.

The expanded covariance matrix for the first three nodes of the example in section 6.3.1 is as follows:

$$\overline{\Sigma_t}=\begin{bmatrix} R_t(1) & 0 & 0 & 0 & R_t(1)a_t(3,1) & 0 & 0 \\ 0 & V_t(1) & 0 & 0 & V_t(1)a_t(3,1) & 0 & 0 \\ 0 & 0 & R_t(2) & 0 & 0 & R_t(2)a_t(3,2) & 0 \\ 0 & 0 & 0 & V_t(2) & 0 & V_t(2)a_t(3,2) & 0 \\ R_t(1)a_t(3,1) & V_t(1)a_t(3,1) & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & R_t(2)a_t(3,2) & V_t(2)a_t(3,2) & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & V_t(3) \end{bmatrix}$$

Unfortunately, covariances between different decomposed parts of the same node (for example $cov(Y_t(3,1), Y_t(3,2))$) are not simple to find. They can, however, be left out of the expanded covariance matrix. The matrix need not be completely populated in order for it to be useful for the forecaster.

In this example:

$$\overline{F_t}=\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Calculating the covariance matrix from the expanded covariance and the above gives:

$$\Sigma_t=\begin{bmatrix} Q_t(1) & 0 & Q_t(1)a_t(3,1) \\ 0 & Q_t(2) & Q_t(2)a_t(3,2) \\ Q_t(1)a_t(3,1) & Q_t(2)a_t(3,2) & \cdot \end{bmatrix}$$

It can be seen that the values missing from the expanded covariance matrix only affect the variances of their corresponding entry in the covariance matrix. This can be corrected easily by calculating the variances through the way shown in section 6.2 and inserting them into the covariance matrix. Once this is done, the covariance matrix is found.

The quantities in the expanded covariance matrix may be of use to the forecaster for informing intervention or to gain a deeper understanding of the individual flows.

6.5 Covariances between entry points

Entry points to the network may present a problem to the calculation of both the covariance matrix and the expanded covariance matrix and thus the production of forecast variances. The covariances between them are not found through the MDM structure and need to be found another way, such as through the sequential technique applied to dynamic Bayes nets in Quintana and West (1987). If an entry point is confounded with other parts of the model the covariance between it and other entry points is not easily found. In the traffic flow application, such a confounded entry point would be a node where traffic flows to a node from one or more parents and also from outside the network. The amount flowing into the network at this node cannot be found directly, which makes calculating the exact covariance between that in-flow and any others impossible. Although Markov Chain Monte Carlo techniques can be used to calculate them in tandem with the regression parameters, this is a lengthy process and might not be practical in a real world setting. In some circumstances, it is possible to assume the covariance between entry points to be zero, in which case the MDM network can model a network of arbitrary complexity, as long as a DAG can be drawn for it. For simplicity, covariances between entry points are assumed to be zero throughout in this thesis, although this is not ideal.

6.6 Overparameterisation as a technique of intervention

This section presents a new technique of intervention. Consider the following form of intervention introduced in section 4.7.2. The modeller intervenes by adding an extra term to the observation equation for one time point:

$$Y_t = F_t^T \theta_t + i_t + v_t$$

where i_t is the quantity by which the modeller is intervening, given by

$$i_t \sim N\left[m_t^*; C_t^*\right]$$

for some values m_t^* and C_t^*. It can be seen that this is functionally identical to adding an additional parameter so that:

$$Y_t = F_t^{*T} \theta_t^* + v_t$$

where:

$$F_t^* = \begin{pmatrix} F_t^T & 1 \end{pmatrix}^T$$
$$\theta_t^* = \begin{pmatrix} \theta_t & i_t \end{pmatrix}^T \sim N\left[\begin{pmatrix} m_t \\ m_t^* \end{pmatrix}; \begin{bmatrix} C_t & 0 \\ 0 & C_t^* \end{bmatrix}\right]$$
$$G^* = \begin{bmatrix} G & 0 \\ 0 & 1 \end{bmatrix}$$
$$\theta_{t+1}^* = G^* \theta_t^0 + W_{t+1}^*$$

This is simply a different representation of acquiring the joint distribution between the intervention parameter and the ordinary parameters (West and Harrison 1999 pg 380). The introduction of the parameter requires a non-square G matrix for one time period, as the additional parameter is added. The system equation is then as specified above until the parameter is removed, again requiring a non-square G matrix. Although the covariances in this equation are given as zero, the model can easily handle non-zero values. The use of zeroes assumes independence between these overparameterisation parameters and the core model parameters. This independent additive assumption need not be made in all cases. Where the intervention is anticipated to occur only briefly in response to a simple event, this assumption may be sensible to make.

This approach provides a posterior for the intervention quantity. The form given in section 4.7.2 does not do this. This can be used to evaluate the accuracy of the expert intervention by considering the difference between the prior and the posterior. This information may then in turn be used to improve expert intervention in the future. In the case where $C_t^*=\infty$ (a 'vague prior'), this approach is identical to discarding the data point. The posterior distributions become the same as the priors.

Further, if the event provoking the inclusion of the intervention parameter i_t is anticipated to occur for several consecutive time periods, i_t can be updated in the same way as the existing parameters in the model. West and Harrison (1999) consider altering a DLM to incorporate an intervention term on a permanent basis- i.e. the event that provoked the intervention has become a usual part of the model. However, there is no reason for the addition to be permanent and the parameter can be removed just as easily, which is not considered in West and Harrison. The i_t term can also be a set of parameters instead of a single value. In this case, the independence between the intervention parameters and the core parameters may not be sensible to assume. However, if the intention is for the intervention parameters to intercept unusual behaviour while the core parameters continue to model the underlying system then independence between them may be assumed initially, although as the model runs the parameters sets will become highly correlated.

DLM models require that the model is of a certain correct structure in order for the forecast variance to converge. At present there is no formal proof of convergence when F_t is not constant in t. By introducing a parameter in this way, the forecast variance may not converge- leading to a breakdown of the model with rounding errors dominating the forecasts. This is known as overparameterisation, and it is possible to reduce an overparameterised model to one that does converge using techniques of model reduction (West and Harrison 1999). In practice the growth of the forecast variance can be slow and the model can continue to run in the short-term without breaking down.

This leads to a further method of intervention here referred to as *overparameterisation* as intervention. This involves introducing additional parameters to the model, then removing them later when intervention is no longer required. This is distinct from a systematic model change (as the intervention is not permanent) and from continuous transient intervention (as the parameters are not generated exogenously at each time point).

Using overparameterisation as intervention is useful when the modeller anticipates a period of unusual activity but does not know what form this activity may take. The added parameters model this deviation from the usual pattern. A consequence of this approach is that the intervention parameters become heavily correlated with the usual parameters. In itself this is not a problem apart from non-convergence- at a certain point rounding errors in the covariances will dominate the forecast. As noted above, for limited periods of use this does not happen. This correlation, however, also has benefits as when the period of intervention ends there is a corresponding increase in forecast variance (removing the need for the modeller to apply this himself to reflect the additional uncertainty surrounding the original parameters now they are no longer confounded with the intervention parameters). If the unusual activity becomes permanent, then existing techniques of model reduction can be applied to reduce the parameter set and regain model convergence.

This is similar to the inclusion of intervention effects (West and Harrison 1999)- but in that case it is assumed that the model remains observable once the intervention parameter has been included. However, in the case of overparameterisation the observability of the model is deliberately broken for a limited period in order to apply continuous intervention that would not otherwise be possible.

Chapter 7 – Applying the MDM to the Traffic Network

This chapter applies the MDM described earlier to the data set introduced in chapter 2. Applying the MDM model to the traffic network requires two particular tasks: determining the structure of the model based on the real world traffic flows and establishing the starting conditions for the model.

7.1 The DAG for the network

The flow diagram (Figure 7.4.1) for the network can be used to heuristically elicit the DAG at time t as described in section 5.2. Due to the missing sites, the non-trivial nodes form two entirely separate graphs. The goal is to construct a DAG including all of these nodes as well as the parameters needed to model them through an MDM.

For this network, the numbering of the counting stations is used as the number for the node (i.e. $Y_t(167)$ is the observed flow at time t for counting station 167). This numbering system does not denote an ordering, although an ordering appropriate for the DAG does exist.

Constructing a DAG heuristically from the flow diagram in figure 2.4.1 is relatively simple. The four nodes $Y_t(167)$, $Y_t(168)$, $Y_t(170a)$ and $Y_t(170b)$ can be represented in a DAG for the system by one root node with three children. Following the principle of putting multiple children in tiers given in section 6.1, this part of the DAG is constructed as shown in figure 7.1.1. In order to differentiate the parameters used for MDM nodes from those used for DLM nodes, the parameters for MDM nodes are given unique Greek letters. In this case α_t represents the proportion of flow

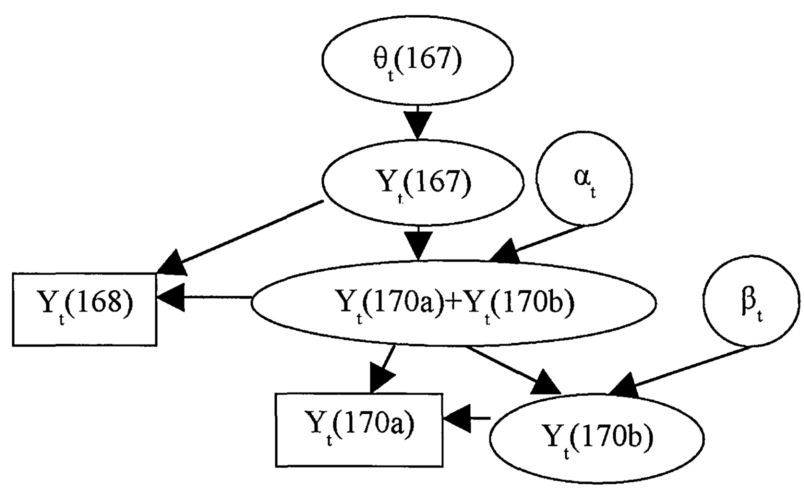

Figure 7.1.1 – Candidate DAG for part of the network

from $Y_t(167)$ that flows to one of $Y_t(170a)$ and $Y_t(170b)$ and β_t represents the proportion of that quantity that flows to $Y_t(170b)$. It is trivial to check that the moralised graph of this subnetwork allows independent updating in the MDM model.

The rest of the DAG can be constructed in a similar way, but there is a complication that means the approach does not work in this case. Where a node has more than one parent, the MDM model requires that the covariance between the parents is known. As shown in section 6, it may not be practical to find this covariance where there are multiple entry points to the network. Covariances between root nodes (or indeed, any entry points to the network) are not easy to find. Without a technique or heuristic to find this covariance, there is little recourse but to assume this covariance is zero. If the actual covariance is positive, as it will be if root nodes follow some global trend, then the MDM will not accurately produce forecast variances and forecast covariances. In practice this leads to a serious problem- an exploding variance in some MDM nodes. The reason behind this is not clear but may be due to the C_t matrix and the estimate of V having to compensate for this bias, or may be due to the discrete nature of the data meaning some observability criterion is broken when the observed values are very low. Without a theory of convergence in non-constant DLMs, it is not possible to say what the cause is with certainty.

What can be done is to combine the parents for a node into a single parent, reducing the number of parameters and thus preventing this between-parent covariance in the parameter set. Adding two parents together in this way again assumes a zero covariance between them, but does not prevent the model from proceeding mechanically. This is not an ideal solution, but it does allow the MDM model to be run for the entire network and can serve as an indicator of how sensitive the MDM is to such problems. Where this problem is exhibited in the network, experimentation was carried out reducing the number of parameters for a given node until the variances converged. Ideally, every in-flow to a node would have its own subset of parameters, whether it is an entry point to the network or a flow from a parent. Instead, all parent flows are summed into a node together with any in-flow. This then serves as the single parent for the node. Both this new summed node and the original node are modelled, but the number of parameters has been reduced. In essence, the one remaining parameter for the node is still a proportion, but the fine detail as to the source of the traffic is lost.

Following this process the second part of the network could be represented by figure 7.1.2. In this case ζ is a parameter modelling $Y_t(163{+}164b)$ as a proportion of $Y_t(162{+}172)$. In the flow diagram (figure 2.4.1) it can be seen that $Y_t(163)$ has an in-flow. This is subsumed into the parameter ζ as part of the parent flow. Because of this, ζ may stray above 1 unlike other regression parameters in the model which should not.

Five quantities are being modelled here: the four flows $Y_t(162) \rightarrow Y_t(163)$, $Y_t(162) \rightarrow Y_t(164b)$, $Y_t(172) \rightarrow Y_t(163)$ and $Y_t(172) \rightarrow Y_t(164b)$ and the in-flow for $Y_t(163)$. These five quantities are modelled through only two parameters leading to a loss of information about the network. This parameter reduction is unfortunately necessary to proceed with the MDM model with the information available. This also breaks the causative relationship between the four observed nodes, which as has been discussed in section 5 is not preferred, particularly when intervention is employed.

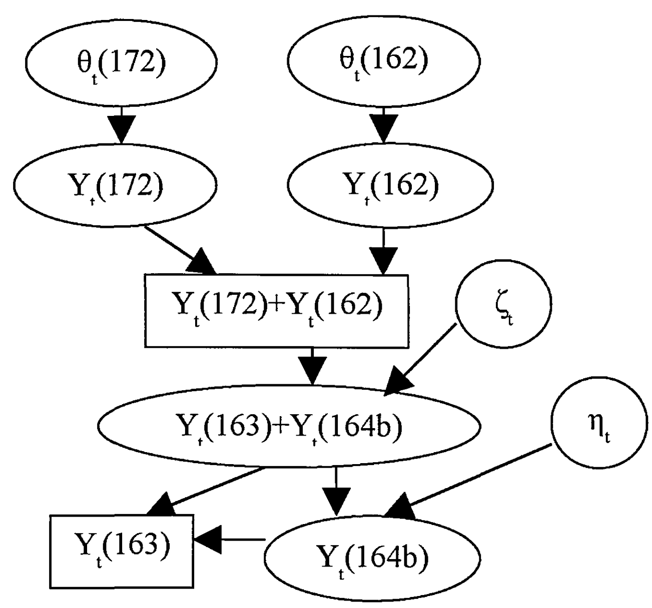

Figure 7.1.2 – Final DAG for the second part of the network

A similar problem occurs in the first part of the network with nodes $Y_t(169)$, $Y_r(170b)$, $Y_t(161)$ and $Y_t(171)$ and is dealt with in the same way. The DAG for the first part of the network is now as shown in figure 7.1.3.

7.2 Leakage

Where traffic from a node only flows to other nodes in the network and not outside the network there is an underlying assumption that the children sum to the parent. This assumption is exploited by deterministic twin nodes and is behind the tiering required with more than two children. The assumption does not necessarily hold as vehicles between counting points when the hour rolls over are counted in the following hour, upsetting this equality. However, in the long term this error has mean zero and will usually be small relative to the total count of vehicles- small enough to be accounted for in the error term. It is advisable to check any such assumptions in the model.

In this network, most nodes have a flow leading outside the network or the assumption is not possible to check as child nodes have a flow of traffic entering the network. The one example where the check is necessary and possible is in the children of node $Y_t(167)$. It has three direct children – nodes $Y_t(168)$, $Y_t(170a)$ and $Y_t(170b)$ – separated out in this DAG in two stages. Examination of the difference between $Y_t(167)$ and the sum of its children reveals that the car count of $Y_t(167)$ is generally larger than the total car count of its children so that some cars are 'lost' between the counting points. This loss is referred to here as 'leakage'. The leakage from $Y_t(167)$ over time is shown in Figure 7.2.1.

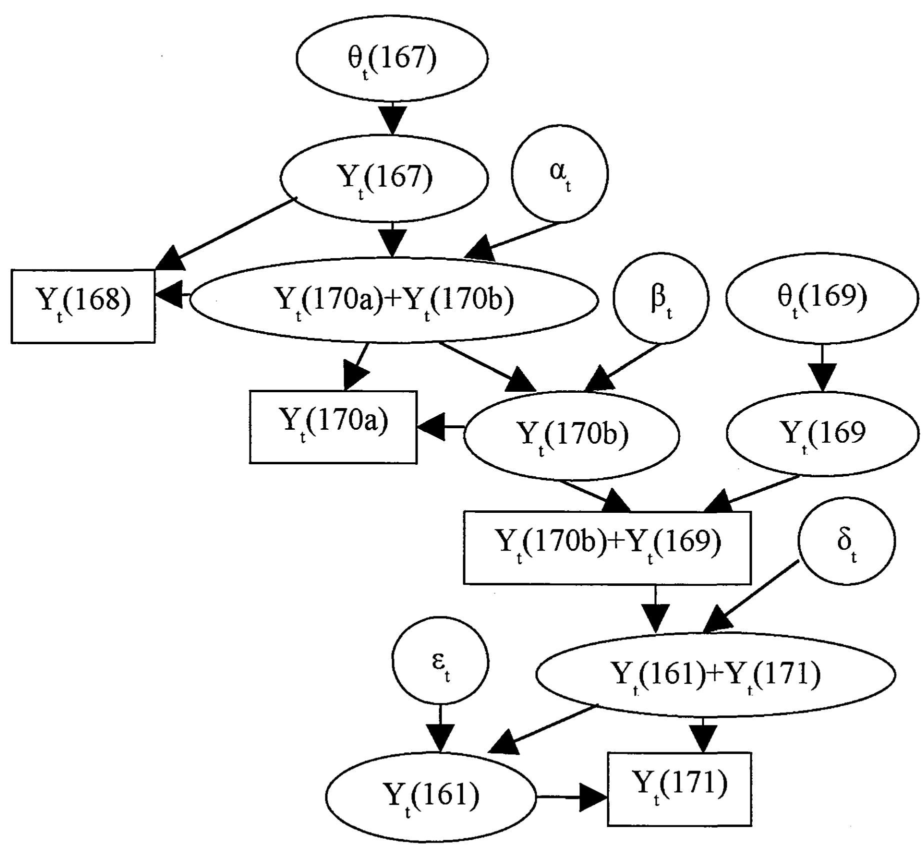

Figure 7.1.3 – Candidate DAG for the first part of the network

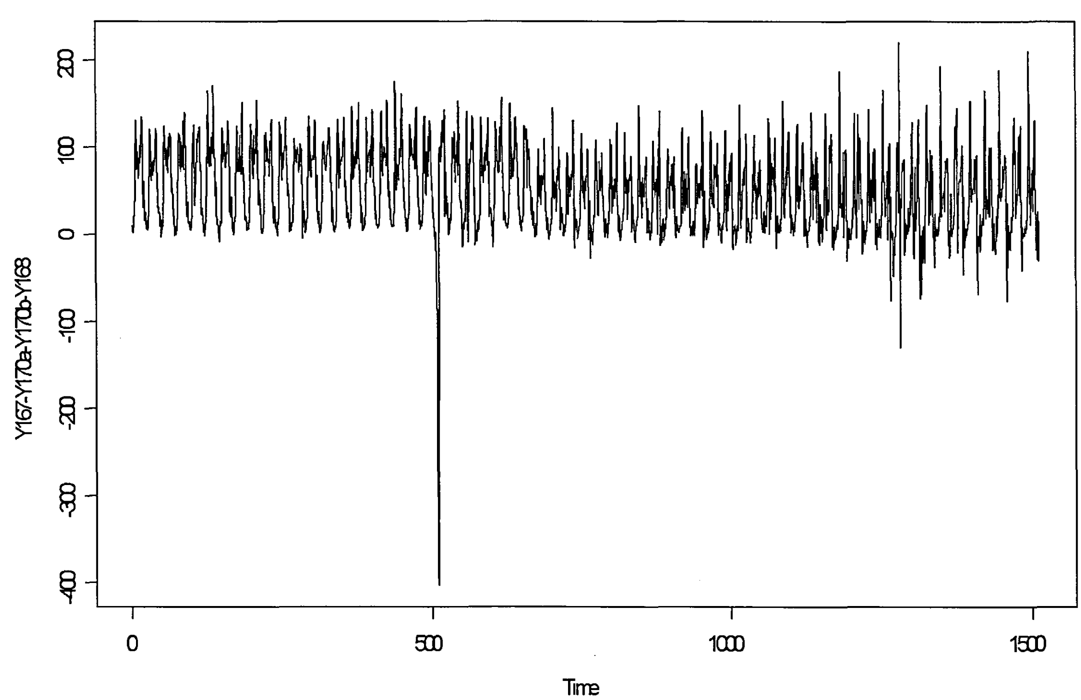

Figure 7.2.1 – Plot of leakage from node 167

Interestingly, the leakage from node $Y_t(167)$ over time follows a seasonal pattern and, more worryingly, it has a non-zero mean. This could be caused by the imperfect nature of the devices used to collect the data. If the four counting points involved have different success rates at counting vehicles, different likelihoods of false positives, or both, then we would expect such a seasonal pattern. Unfortunately, the leakage pattern above does not yield itself to such a correction. It is possible to use maximum likelihood to find the linear combination that best fits the above data, but even this is a poor fit. It would appear that the mechanism behind the inaccurate counting is complex and not readily discerned given the data available. The providers of the data (Babtie Reading Highways Agency Traffic Team) suggested that exogenous factors such as vehicle speed or road surface conditions may be influencing the value. A technique for accounting for measurement error with this kind of counting station is given in Hazelton (2001), but for the purposes of this model, this leakage is most elegantly modelled by considering it as a separate quantity. It is denoted by $Y_t(L)$ on the DAG for the network, and the quantity $Y_t(168L)$ is the sum of this quantity and traffic flow past node $Y_t(168)$. This final DAG for the first part of the network is shown in figure 7.2.2. The parameter γ_t is the proportion of $Y_t(168L)$ which is observed in $Y_t(168)$.

7.3 Modelling choices

7.3.1 Choice of DAG

While the DAGs produced above are appropriate for the model, there is not a single unique DAG that encapsulates the network. The DAGs above closely match

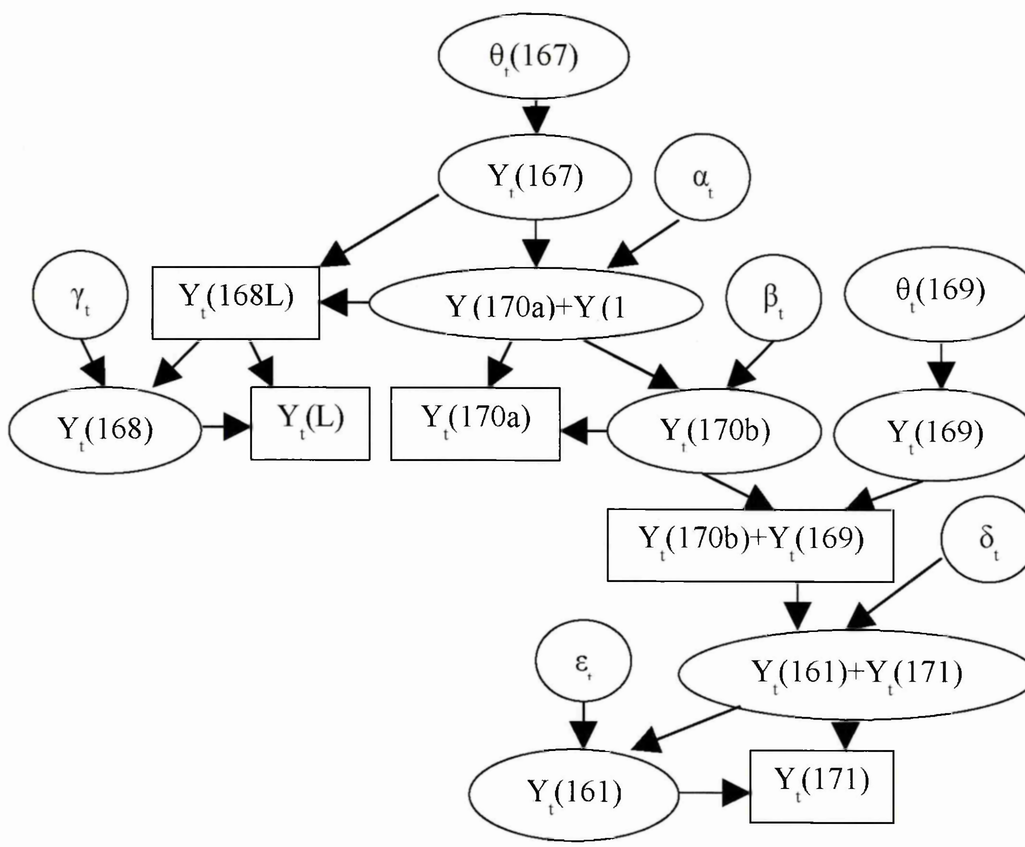

Figure 7.2.2 – Final DAG for the first part of the network.

the actually flow of traffic through the network and are drawn to make the model as easy to follow and as interpretable as possible. The parameters are also easy to interpret. Where compromises have been made in order to apply the MDM, this will be a good indicator of how sensitive the MDM is to such potential problems. These DAGs given in figures 7.1.3 and 7.2.2 are used for the model henceforth.

7.3.2 Choice of model for entry points

Where an MDM node has both parents and in-flows from outside the network, the model structure requires that a DLM be used to model this inflowing traffic as shown in section 5.2. This is part of the MDM model and attracts no additional complication. However, there are no longer any such nodes in the network due to the parameter reduction performed in constructing the DAG. Where a node is an entry point and has no parents there is greater leeway in what method can be used to model it; in essence any methodology could be used. For this model we will use standard DLM models for such entry points for convenience and to gain the advantages of the DLM-based approach.

7.3.3 Choice of model for seasonality

Of the three principle means of modelling seasonality introduced in section 4.6, seasonal factors are used. They are preferred to Fourier models as they are easier to interpret and thus simplify intervention. Full-form Fourier models give identical mean squared errors, and reduced form Fourier models perform less well. The seasonal effects approach is not used as there is no sensible typical value to estimate. These seasonal factor models are used for all nodes in the model. Thus, for a DLM root node:

$$F=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix};\qquad G=\begin{bmatrix}0&1&0&\cdots&0\\\vdots&\ddots&\ddots&\ddots&\vdots\\\vdots&&\ddots&\ddots&0\\0&&&\ddots&1\\1&0&\cdots&\cdots&0\end{bmatrix}$$

For an MDM node with one parent:

$$F_t(i)=\begin{pmatrix}y_t(j)\\0\\\vdots\\0\end{pmatrix};G=\begin{bmatrix}0&1&0&\cdots&0\\\vdots&\ddots&\ddots&\ddots&\vdots\\\vdots&&\ddots&\ddots&0\\0&&&\ddots&1\\1&0&\cdots&\cdots&0\end{bmatrix}$$

where j is the parent of i. For MDM nodes with more than one parent or an in-flow, the superposition principle could be applied but the parameter reduction has eliminated any such nodes from the model.

7.3.4 Choice of priors

In a time series of this length the choice of priors should not be critical to the model's long term performance. To improve early performance of the model informative priors are used. The first week of data is used as a training set to estimate the point forecast and observation noise variance for each node. The system noise variance is selected as a deliberately high value to allow the model to converge rapidly on an accurate value.

7.3.5 Other model choices

A selection of DLM and MDM nodes were run using different discount factors to establish which gave the best performance of the model. On the basis of mean squared error, the discount factor of 0.98 was chosen for both types of node. This value was adopted throughout the model.

7.3.6 Final models chosen

The final models for the network were as follows. The root nodes in the network were modelled as DLM models of the following form:

$$Y_t(i)=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}_{24\times1}^{T}\theta_t(i)_{24\times1}+v_t(i) \qquad v_t(i)\sim N\left[0;V(i)\right]$$

$$\theta_t(i)=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \ddots & 0\\ 0 & & & \ddots & 1\\ 1 & 0 & \cdots & \cdots & 0\end{bmatrix}_{24\times24}\theta_{t-1}(i)_{24\times1}+w_t(i) \qquad w_t(i)\sim T_{n_t-1}\left[0;W_t(i)\right]$$

with discounted $W_t(i)$ and variance learning for V as given in section 4.5.

The models for MDM nodes, all of which have one parent in this DAG, are of the following form:

$$Y_t(i)=\begin{pmatrix}pa(Y(i))_t\\0\\\vdots\\0\end{pmatrix}_{24\times1}^{T}\theta_t(i)_{24\times1}+v_t(i) \qquad v_t(i)\sim N\left[0;V(i)\right]$$

$$\theta_t(i)=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & & \ddots & \ddots & 0\\ 0 & & & \ddots & 1\\ 1 & 0 & \cdots & \cdots & 0\end{bmatrix}_{24\times24}\theta_{t-1}(i)_{24\times1}+w_t(i) \qquad w_t(i)\sim T_{n_t-1}\left[0;W_t(i)\right]$$

again with variance learning and discounting.

7.4 Model performance

When inspecting plots of the forecast errors for the model, periods of unusual activity can be identified informally by inspecting the forecast errors and variances.

Periods of unusual activity, even when very short, are signalled by increased errors and increased forecast variances in following time periods. For the node $Y_t(170b)$, for example, we can plot the forecast errors in Figure 7.4.1. The solid line denotes the one-step ahead forecast errors, and the dotted line is the square root of the one-step ahead forecast variance. It is possible to see how periods of unusual activity tend to increase the forecast variance, as evidenced by the behaviour of the model from time 750 to 950. The seasonality of the variance (discussed in section 6.1) is also noticeable.

The Mean Squared Error is used as the primary means of comparing performance between models, and the Median Forecast Variance is an informal method of comparing how precise models are in their forecasts. The first section of table 7.4.1 contains root nodes that are DLMs, the rest of the table are MDM nodes.

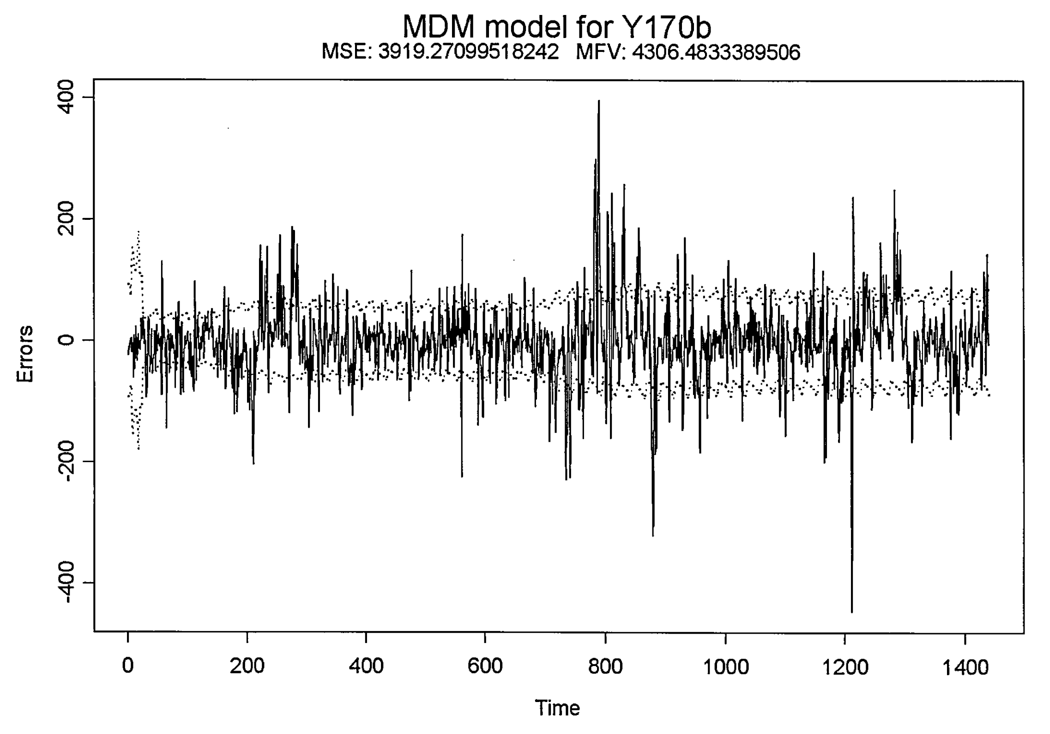

Figure 7.4.1 – Forecasts errors and forecast variances for node Y170b in the MDM model

Node	MSE	MFV
162	4374	4190
167	41948	37989
169	678	641
172	7387	7225
161	11299	24662
163	937	1126
164b	6244	4044
168	902	1645
170a	17122	10810
170b	5425	4327
171	43967	26819

Table 7.4.1 – Performance of the MDM network.

It can be seen that the MFV is similar to the MSE for many nodes which is only to be expected when modelling data expected to exhibit Poisson behaviour.

Examining the autocorrelation of the forecast errors gives Figure 7.4.2. These plots show there is some autocorrelation at lag 1 unexplained by the model, but modelling this would require either the use of a Fourier form model or a significant change in the structure of the model. Either of these would complicate the model to the point where its key benefits (in terms of interpretable parameters, short calculation time and easy intervention) would be lost. Other slight correlations could be the result of evening rush hour traffic flow following morning rush hour traffic flow, but without further contextual data it is difficult to draw strong conclusions.

Checking the distribution of the errors yields Figure 7.4.3. The distribution of the errors is not particularly normal. The errors seem to be heavy tailed, which is

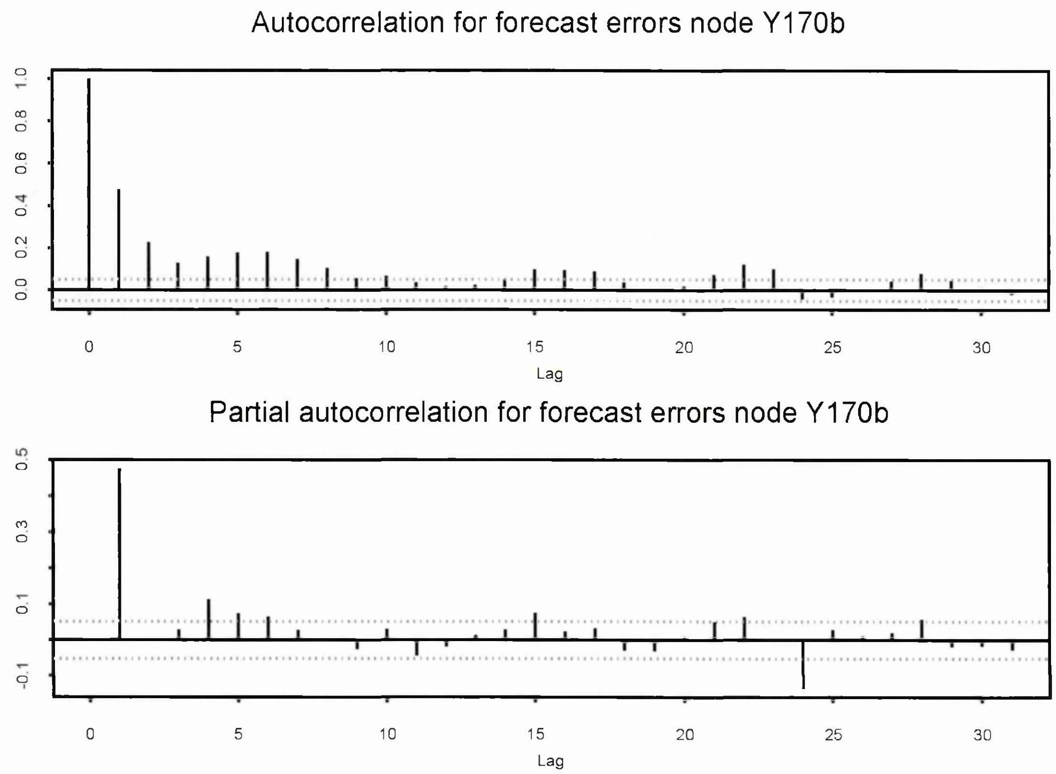

Figure 7.4.2 – Autocorrelation of forecast errors for node Y170b in the MDM model

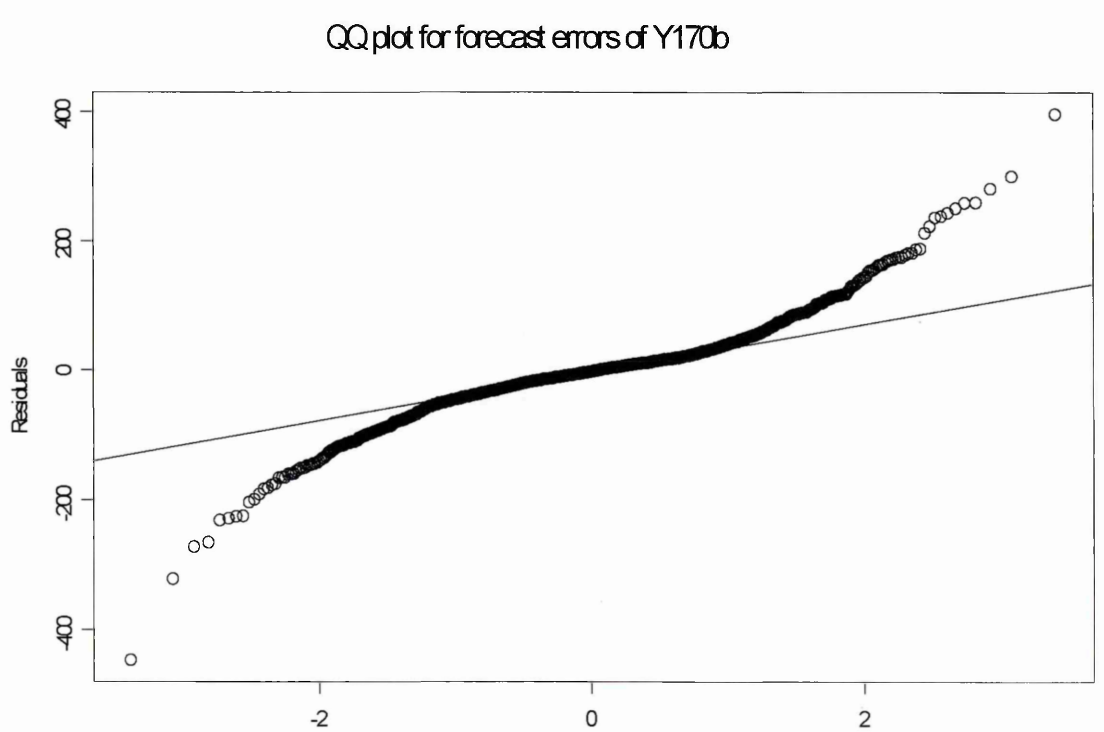

Figure 7.4.3 – Residuals plot of the forecast errors for node Y170b in the MDM model.

what would be expected from a model using T-distributions. Equivalent plots for other nodes yield similar results. Heavy tails might also be expected in a data set with periods of unusual activity. In order to improve the performance of the model, these periods of unusual activity will have to be accounted for.

7.5 Intervention

Expert intervention can be introduced into the model in order to improve its performance. Where there is bad weather, roadworks, accidents or congestion, intervention can model the changes to the usual pattern easily as the MDM uses univariate DLMs. Intervention requires an expert to assess the data in real-time and make decisions regarding whether intervention is necessary and what form it should take. The expert need not be the forecaster implementing the model, although the forecaster will then need to interpret the expert's information in order to apply it. As motorways are monitored in real-time already, at least during the day time, such an approach is practical. Intervening for simple events such as those described above requires no structural changes to the model. Where a change is more serious (such as one road blocked completely or a new stretch of road built) the model must change structurally. However, the existing model can be used to construct the new model, including using existing parameters to generate priors for the new model. Indeed, much of the model may be unaffected after such a structural change. For example consider figure 7.2.2. If a new node was introduced after node Y(169) which monitored the traffic leaving the network after that node then the structure for nodes Y(161) and Y(171), which are below node Y(169), would change. The rest of the network, including the nodes Y(167), Y(168), Y(170a) and Y(170b), would be unaffected. The priors for this new model would be based on the parameters from the old model. Thus, much of the information learned about the network before the change would still be used in the revised model.

In a large network, because of the highly multivariate nature of the problem, intervention usually needs to affect a large number of nodes when an unusual event occurs. The MDM model can reduce this amount of work considerably as the hierarchical structure of the MDM model allows intervention to percolate downwards through the model from where it takes place (as shown in section 5.4.1).

The data are supplied without context which makes legitimate intervention difficult. In a real-time forecasting system the modeller would have access to good information regarding the traffic conditions, for example, pictures of current motorway conditions and notification of any future events that may change traffic flows. Without an outside source of information, all intervention must be conducted on the basis of the raw data. If the modeller uses future data points to inform his intervention, then care must be taken not to be too prescient. Not all events that might prompt intervention need be due to exogenous factors- the capacity drop associated with congestion has several theories that reproduce it (for a recent example see Zhang and Kim, 2005) based on the traffic data itself. This raises the possibility that automated event detection is feasible and the MDM model can incorporate such endogenous information in the form of intervention.

When modelling a network such as this, unusual events can be categorised as short-term transient, long-term transient and permanent. The traffic flow returns to its original pattern after transient events. In this thesis events are classified as short-term transient events if they last for a day or less and long-term transient events if they last for longer than a day. The distinction between short-term and long-term transient events is a convenience and is by no means significant or universal. In these data there are no permanent events identifiable.

The assumption used for the following intervention is that short-term transient events have no prior warning, but the modeller knows when long-term transient events will occur and what form they are likely to take. For short-term events the assumption has been made that the modeller can only intervene for the first time point by disregarding the data point (treating it as an outlier or missing data). This allows the modeller to avoid biasing the parameters but does not improve the forecast error for that point. For subsequent time points in short-term transient events and all time points in long-term transient events, the modeller can intervene as they see fit. The intervention performed under these assumptions should provide a credible assessment of how intervention would affect model performance if exogenous information were available.

In a model such as this, there are likely to be characteristic patterns of unusual activity for common events like congestion, accidents and special events in the immediate area. The modeller can have 'stock responses' for such events to make intervention simpler still. Maunch and Cassidy (2002) demonstrate that a traffic queue for a given location behaves in a particular way and justifies such an approach. A technique of traffic forecasting that used periods of historical data as templates for current conditions instead of time series methodology was explored in Wild (1997). A variant of that technique could be used in parallel with the methodology described here to handle unusual (but irregularly recurring) events.

7.5.1 Intervention locations

Locations in the data where intervention is applied are identified informally from inspection of the data and forecast errors without intervention. This is information that the modeller in a real-time application of this model would have available when the event is unanticipated. The nature of the intervention and its extent are informally decided based on the assumptions in the section 6.5. Intervention could be done formally with a monitoring system- but ultimately the form of the monitoring system would depend heavily on the purpose of the forecast. As yet, no formal monitoring system has been developed for this model. Intervention is kept informal here for simplicity. Applying formal intervention in an MDM model of this data set is a possibility for further work.

The instances of intervention are listed tier by tier, as described in section 5.4.1. The figures 7.1.2 and 7.2.2 show the DAG and the tiers used here. Although the two sections are completely separate and could be considered simultaneously, here the smaller part shown in 7.1.2 is assumed to be after that in 7.2.2. Tiers 1 to 6 are in the larger part of the network, tiers 7 to 10 are in the smaller.

Tier 1- node Y(167)

This node is a simple DLM. The first event occurs at time 488 in figure 7.5.1.1. This is a depressed flow followed by an inflated flow. This could be characteristic of a period of congestion- traffic grinds to a halt in the first time period (reducing the flow), then once congestion is eased the flow is inflated as the congested cars clear. It is short-term transient. The first time point can be considered an outlier and the second can be adjusted for. For congestion, a sensible choice of

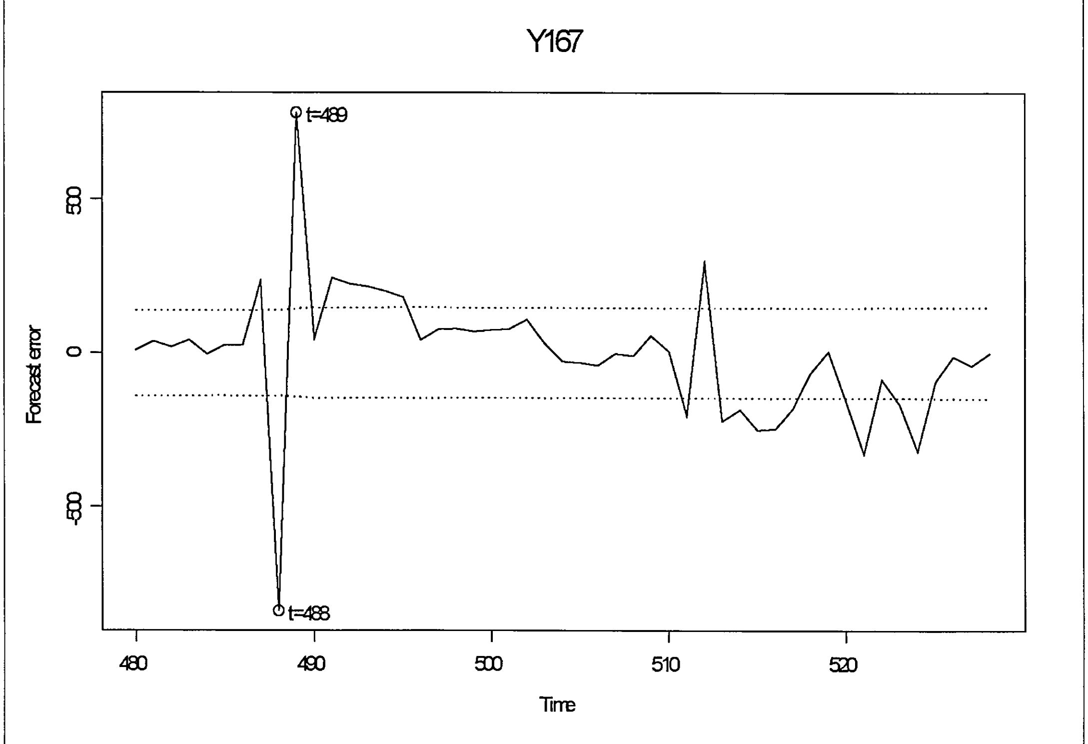

Figure 7.5.1.1 – Section of forecast errors for Y167

intervention parameter is of the same magnitude as the error in the first time point- the cars that couldn't get through in the previous hour do so now.

Time	Type of intervention	Intervention value	Intervention variance
488	Disregard data point		
489	Add parameter for one time point	700	10000

Table 7.5.1.1 – First intervention performed at node Y(167)

The result of this intervention is shown in figure 7.5.1.2. The forecast error for the second time period of note is much smaller. The increase in forecast variance can be seen not just for that time point, but 24 hours later when the model uses that parameter again.

The second event occurs from time 625 to 792 in figure 7.5.1.3. This is a period of unusual activity where the lunchtime traffic flows are unusually high, but not in a way that lends itself to a simple explanation. As a long-term transient event, it is assumed that the modeller has prior knowledge that the event will occur, but its peculiar nature suggests that the modeller may not have a simple intervention to perform. The poor model performance after time 800 may be a result of the unusual activity earlier biasing the parameters. This is a good place to attempt to apply the overparameterisation intervention technique described in section 6.6- and indeed was the motivation for its formulation. The unusual activity is not of a simple form- there is an increase in flows during certain hours of the day but not others, and the amount by which it changes follows a unimodal curve. The complicated nature of the event suggests that even if the forecaster knew the event was going to occur they would not

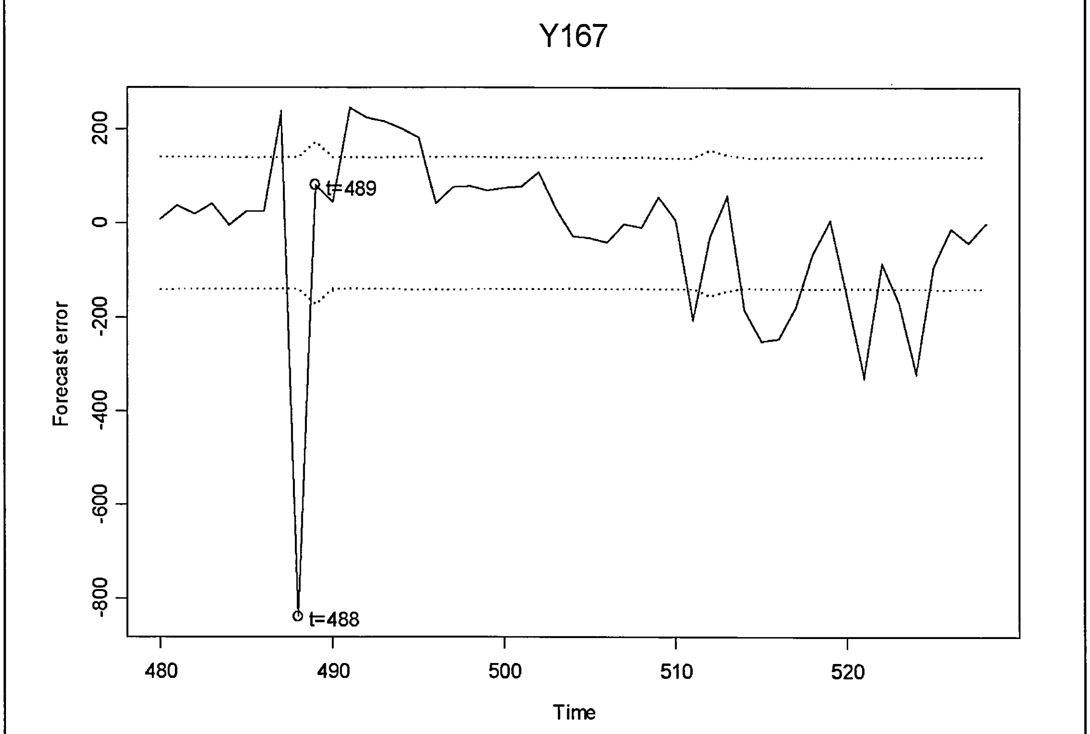

Figure 7.5.1.2 – Section of forecast errors for Y167 after intervention

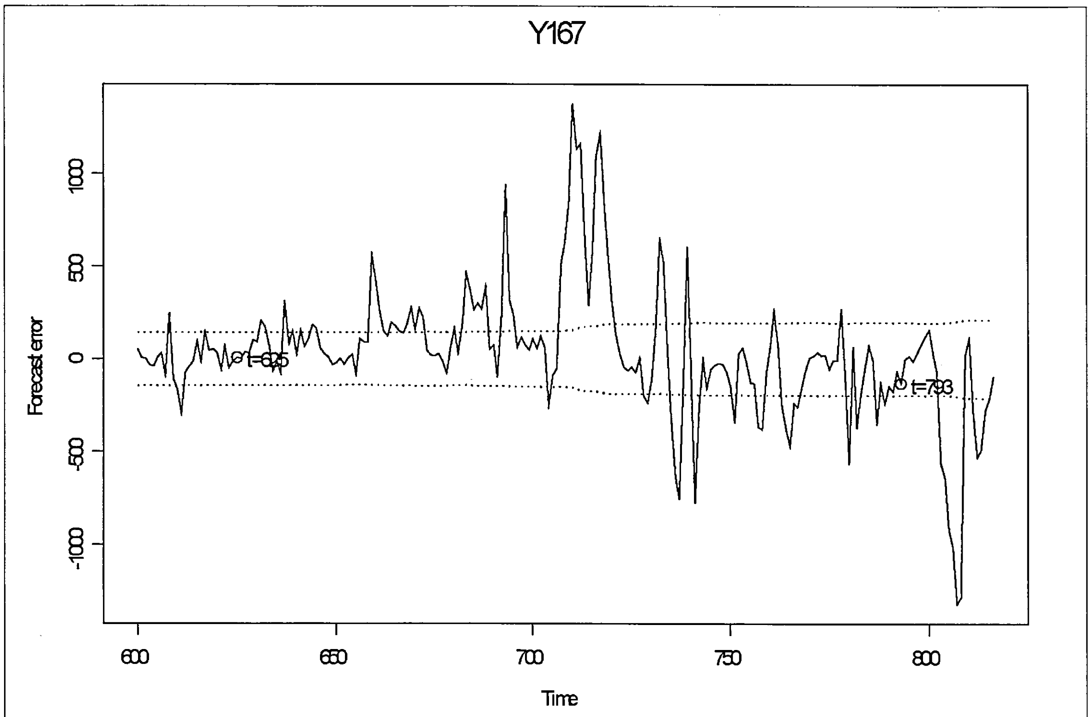

Figure 7.5.1.3 – Section of forecast errors for node Y167

know by what amount to intervene. This suggests introducing overparameterisation with parameters set to zero that are intended to track this pattern

Time	Type of intervention	Intervention value	Intervention variance
625	Add overparameterisation parameters to model	$(0 \cdots 0)^T$	50000×I
793	Remove overparameterisation parameters from model		

Table 7.5.1.2 – Second intervention performed at node Y(167)

A large intervention variance is specified to allow the parameters to adapt quickly. The behaviour of the model when this intervention is performed is shown in figure 7.5.1.4.

There is no visible improvement in performance, although examination of the mean squared error over the time period in question does show a significant improvement. The activity is sufficiently chaotic to suggest that visible improvement is unlikely in any realistic circumstance. The additional forecast variance when the parameters are applied is easy to identify. The next day shows that the forecast variance has returned to a level similar to that before the intervention. During the intervention the forecast variance grows- although it seems that the extreme values are more significant in its rise than the overparameterisation. After the parameters are removed, there is another rise in forecast variance for a day before it returns to usual

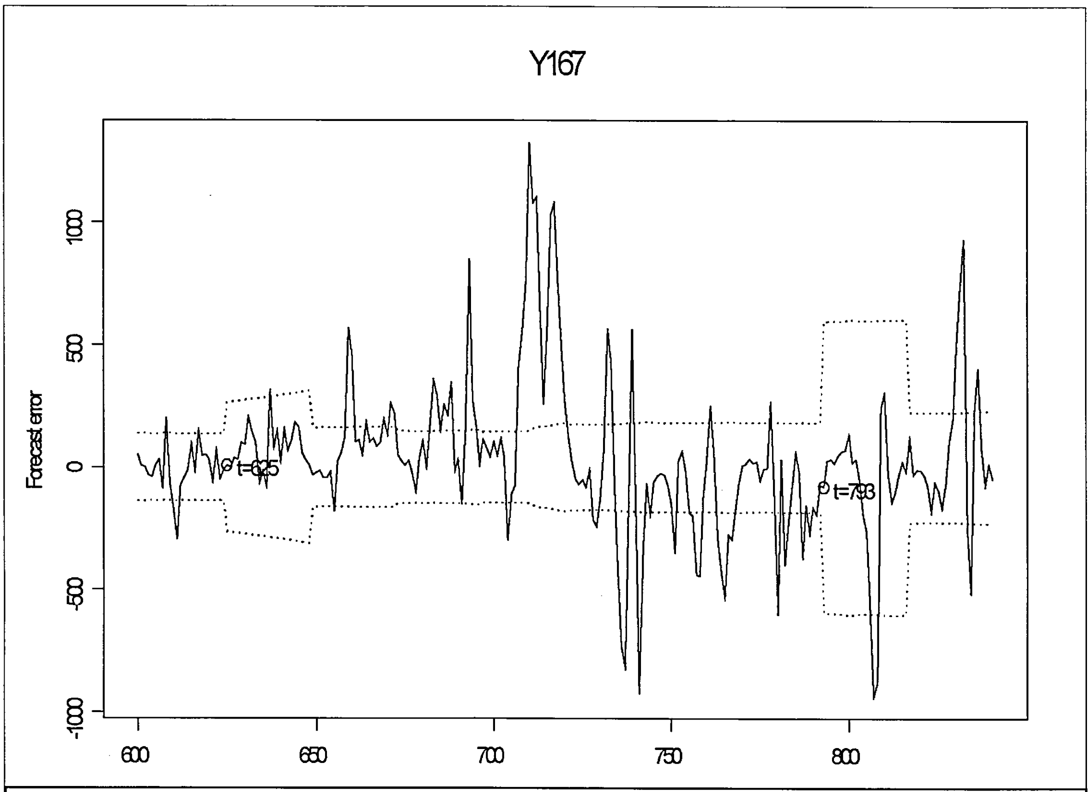

Figure 7.5.1.4 – Section of forecast errors for node Y167 after intervention

levels. The intervention parameters at the end of the intervention were non-zero, still with very high variances associated with them. There was also a strong negative covariance between the intervention parameters and the regular parameters, as would be expected.

The final event for this node occurs from time 1139 to 1143 in figure 7.5.1.5. This looks like a prolonged period of congestion, with a single time period in the middle where flow is normal as the queue begins to clear. The modeller is assumed to know how long this congestion will continue for- a realistic assumption as a real-time traffic monitoring system would most likely have cameras in place to watch current road and traffic conditions. This first time point is an outlier (as the event is short-term transient), and the subsequent time points can be adjusted for appropriately.

Time	Type of intervention	Intervention value	Intervention variance
1139	Disregard data point		
1140	Add parameter for one time point	-800	10000
1142	Add parameter for one time point	800	10000
1143	Add parameter for one time point	800	10000

Table 7.5.1.3 – Third intervention performed at node Y(167)

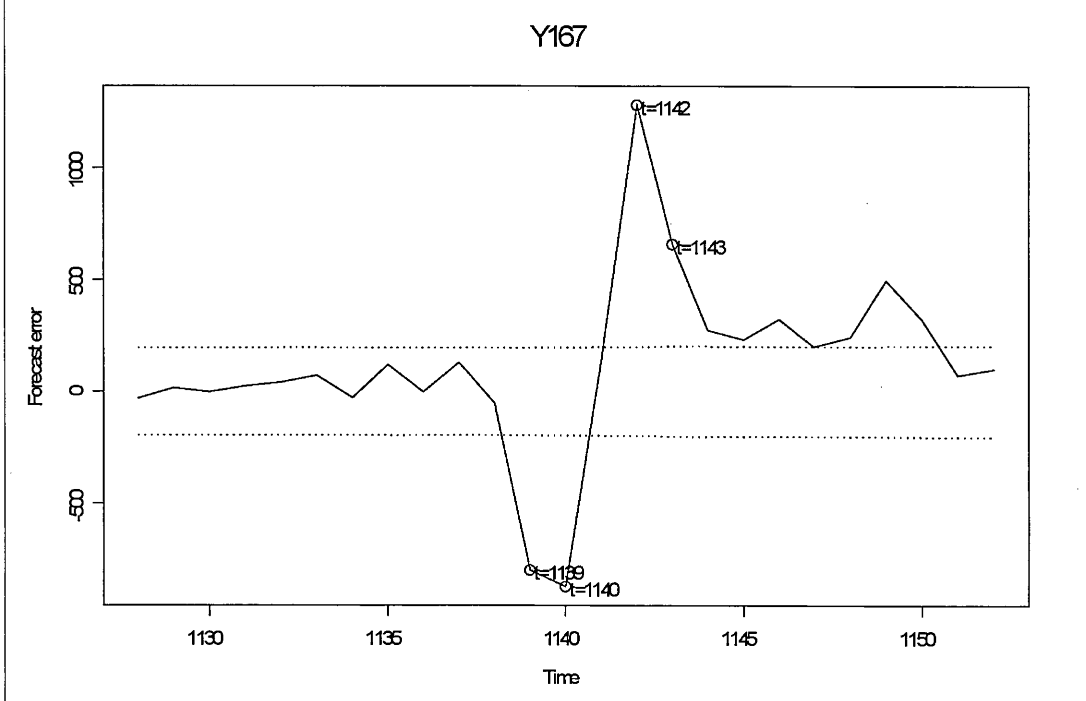

Figure 7.5.1.5 – Section of forecast errors for node Y167

The values for the intervention were found by examining the forecast errors of the nodes in the event. Without the benefit of hindsight, the expert would have to arrive at such figures a different way. The impact of intervention is shown in figure 7.5.1.6. As in the first intervention, there is improvement in the forecast error.

Tier 2 – nodes Y(168L), Y(170a)+Y(170b)

Following this intervention, no further intervention is required for nodes Y(168L) or Y(170a)+Y(170b) as the intervention in node Y(167) percolates through to them. No intervention is required at tier 2.

Tier 3 - nodes Y(168), Y(L), Y(170a), Y(170b), Y(169)

For node Y(168), there is one highly unusual event in this time series starting at time 437 in figure 7.5.1.7. This pattern is very different from the usual activity for this node. Not only is the magnitude far higher than the rest of the errors, but examination of the raw data shows that this activity is a result of a huge number of vehicles appearing between node Y(167) and node Y(168). It is possible that the event is the result of faulty recording equipment. There is not a consistent value for these errors and the pattern is not repeated, so without expert information it is most sensible to treat them as pure outliers. This is summarised in table 7.5.1.4.

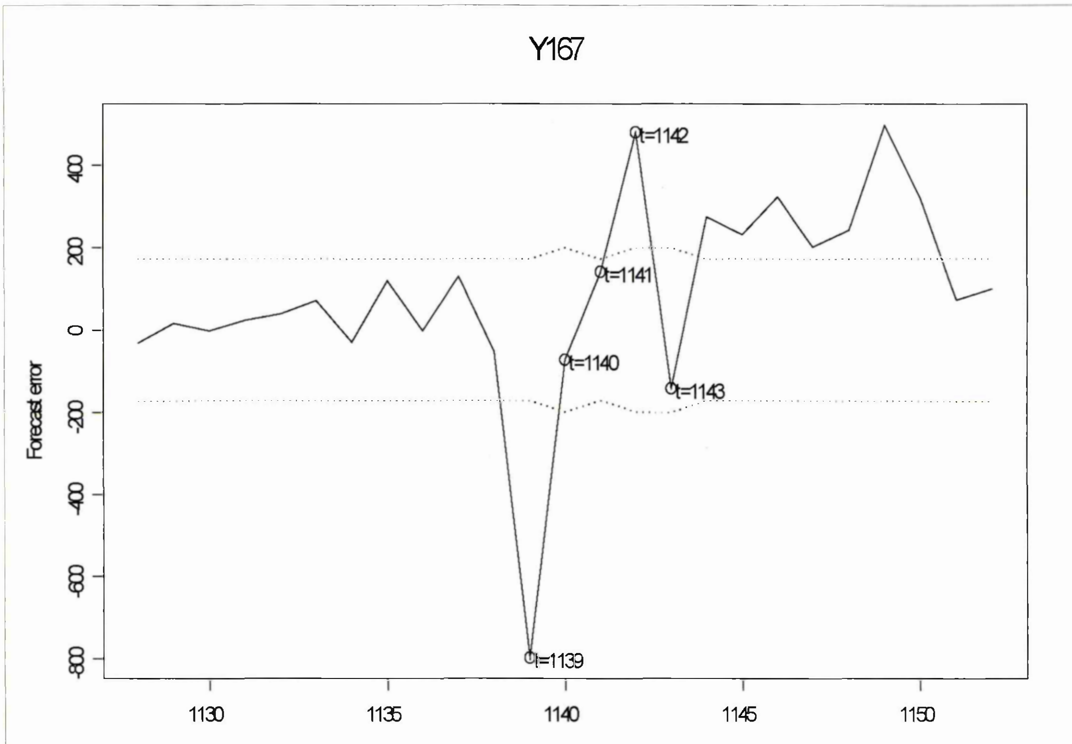

Figure 7.5.1.6 – Section of forecast errors for node Y167 after intervention

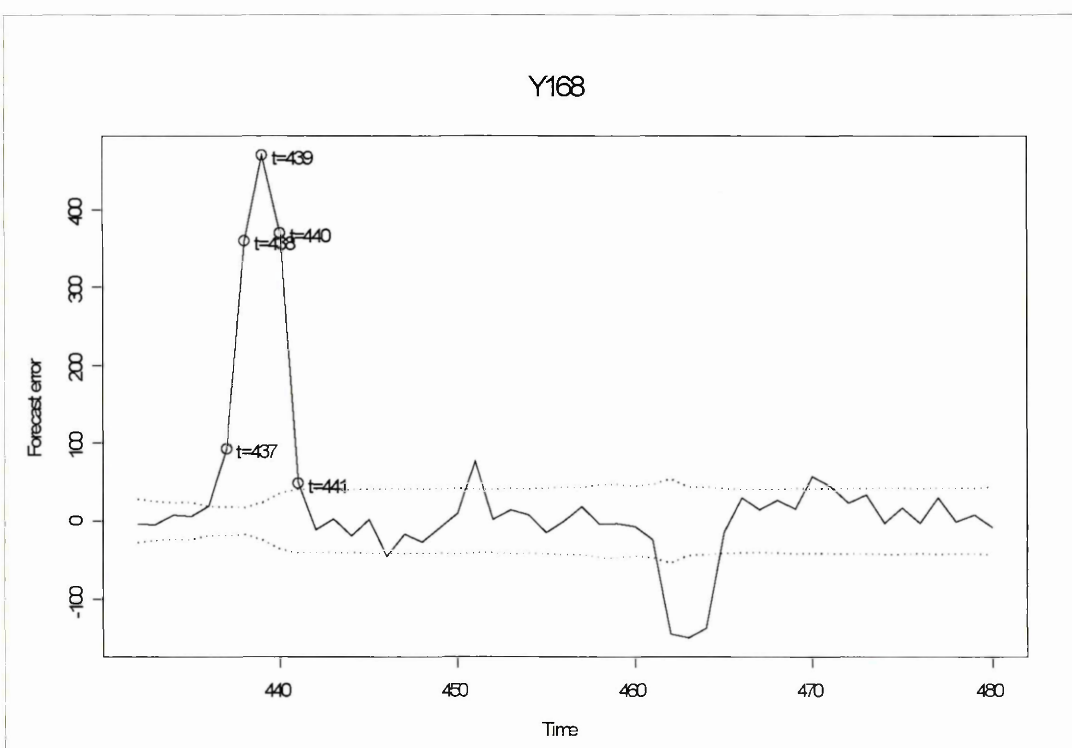

Figure 7.5.1.7 – Section of forecast errors for node Y168 after intervention at node Y167

Time	Type of intervention	Intervention value	Intervention variance
437	Disregard data point		
438	Disregard data point		
439	Disregard data point		
440	Disregard data point		
441	Disregard data point		

Table 7.5.1.4 – Intervention performed at node Y(168)

The result of the intervention is given in figure 7.5.1.8. Although the intervention does not improve the forecasts during the period, it improves them one day later. The outliers biased the parameters leading to poor forecasting in the next cycle. By treating the data points during the event as outliers, the parameters do not become biased.

Node Y(L) is not a quantity of particular interest, and in any case is determined by the model for Y(168). Deterministic twins should not have intervention per se, intervention should be performed on their sibling. The remaining nodes in this tier do not require any intervention.

Tier 4 – node Y(170b)+Y(169)

This node is deterministic and the sum of two other nodes. There is no need for intervention in it.

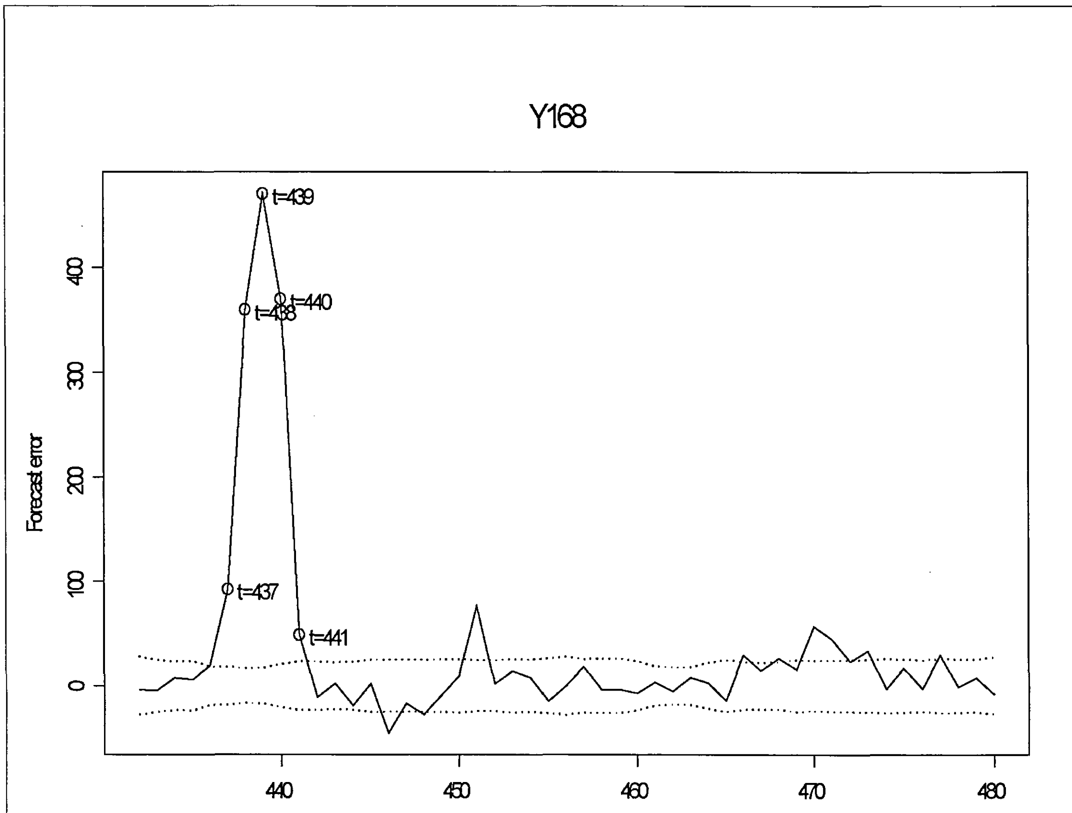

Figure 7.5.1.8 – Section of forecast errors for node Y168 after full intervention

Tier 5 – node Y(161)+Y(171)

This node was created in order to have a valid DAG for the MDM model. However, intervention in this node is complicated as any unusual activity is a result of unusual activity in one, or both, of the nodes 161 and 171. If, for example, the activity here stems from activity in node 171, then any intervention here will erroneously percolate to node 161, requiring further intervention there. This is because the DAG selected does not reflect the causal relationships between the nodes correctly, as described in section 5.2.

However, intervention can still take place.

The first event for node Y(161)+Y(171) begins at time 265 in figure 7.5.1.9. However, by examining the observed values of Y161 and Y171 it can be seen that the event is caused by a drop in flow for Y171 seen in figure 7.5.1.10. The flow appears to be approximately half the usual during this period. This translates to a proportion of 17/20 in node 161+171 for this long-term transient event. At the start of the period of intervention, all the δ_t parameters (from figure 7.2.2) are multiplied by the proportion 17/20. The variance for the parameter estimate is increased at the same time. As this node is an MDM node the model parameter is a regression parameter so this increase is on the scale of proportions and not absolute values. It is assumed that once the activity ends the inverse proportion is a suitable means of restoring the regression parameters to their previous level. When the period ends the parameter set is multiplied by 20/17 in the same way.

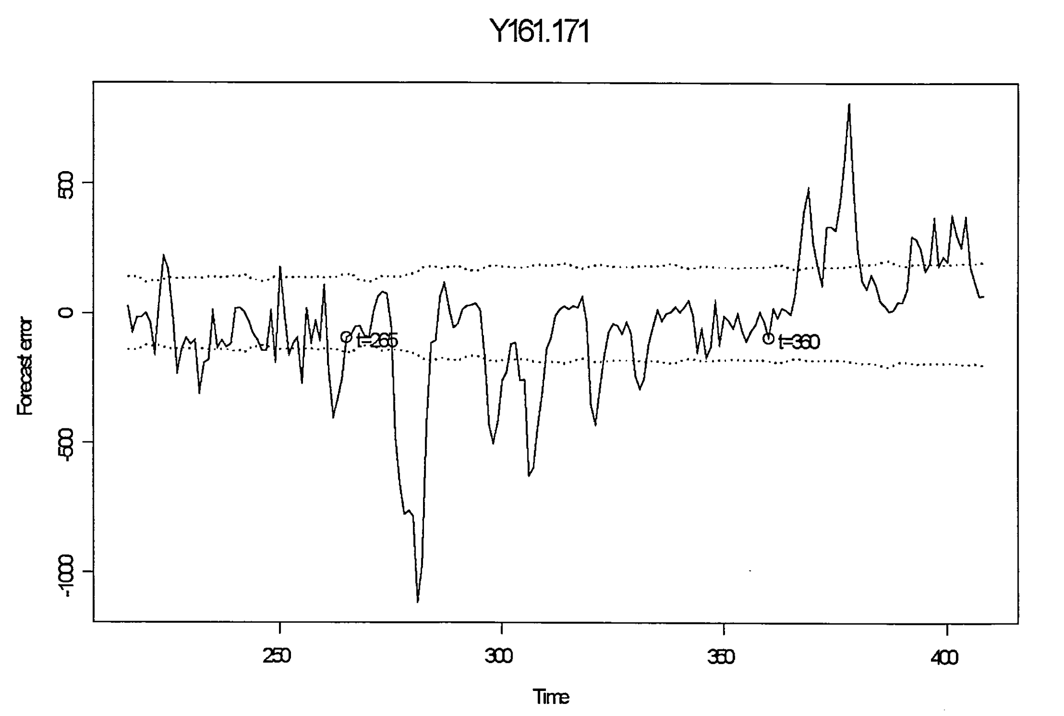

Figure 7.5.1.9 – Section of forecast errors for node Y161+171 after intervention at node Y167

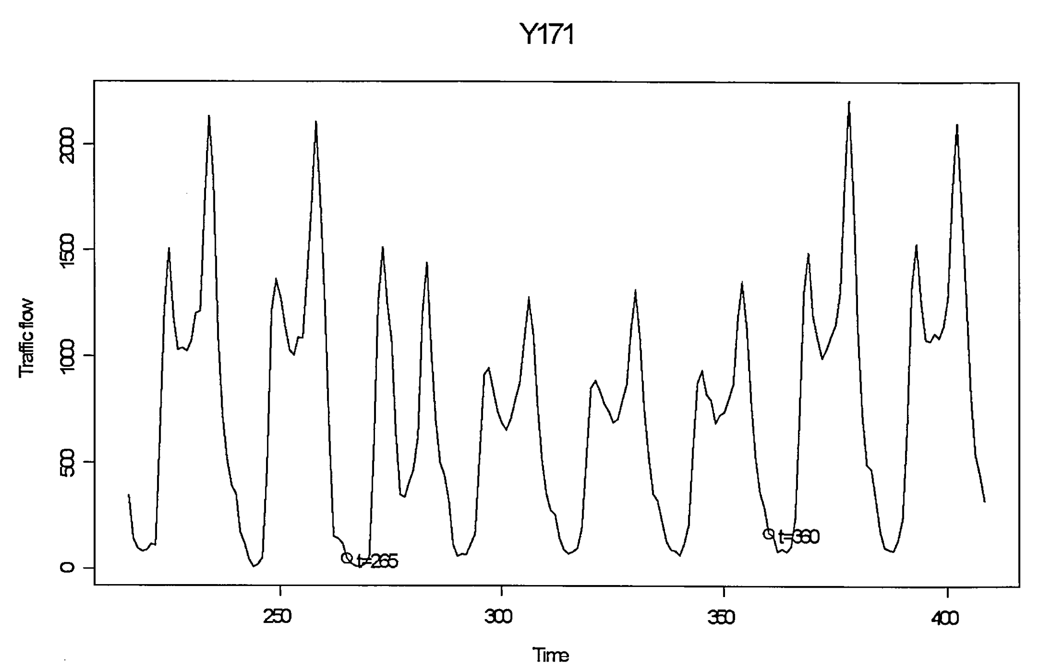

Figure 7.5.1.10 – Section of raw data for node Y171

Time	Type of intervention	Intervention value	Intervention variance
265	Scale δ_t by a value	×17/20	+0.01
360	Scale δ_t by a value	×20/17	+0.01

Table 7.5.1.5 – First intervention performed at node Y(161)+Y(171)

The affect of the intervention can be seen in figure 7.5.1.11. The forecast errors are still large, but are more symmetrical and are more within the error bounds from the increased variance.

The second event begins at time 433. The behaviour of this event is the same as for the previous one, even down to the proportion, so the same intervention method is used.

Time	Type of intervention	Intervention value	Intervention variance
433	Scale δ_t by a value	×17/20	+0.01
576	Scale δ_t by a value	×20/17	+0.01

Table 7.5.1.6 – Second intervention performed at node Y(161)+Y(171)

The final event begins at time 1068 in figure 7.5.1.12. There is a period of depressed values, seeming to progress in steps. Although the first step is unusual, it is not dramatically so and it is reasonable to assume the modeller would not intervene

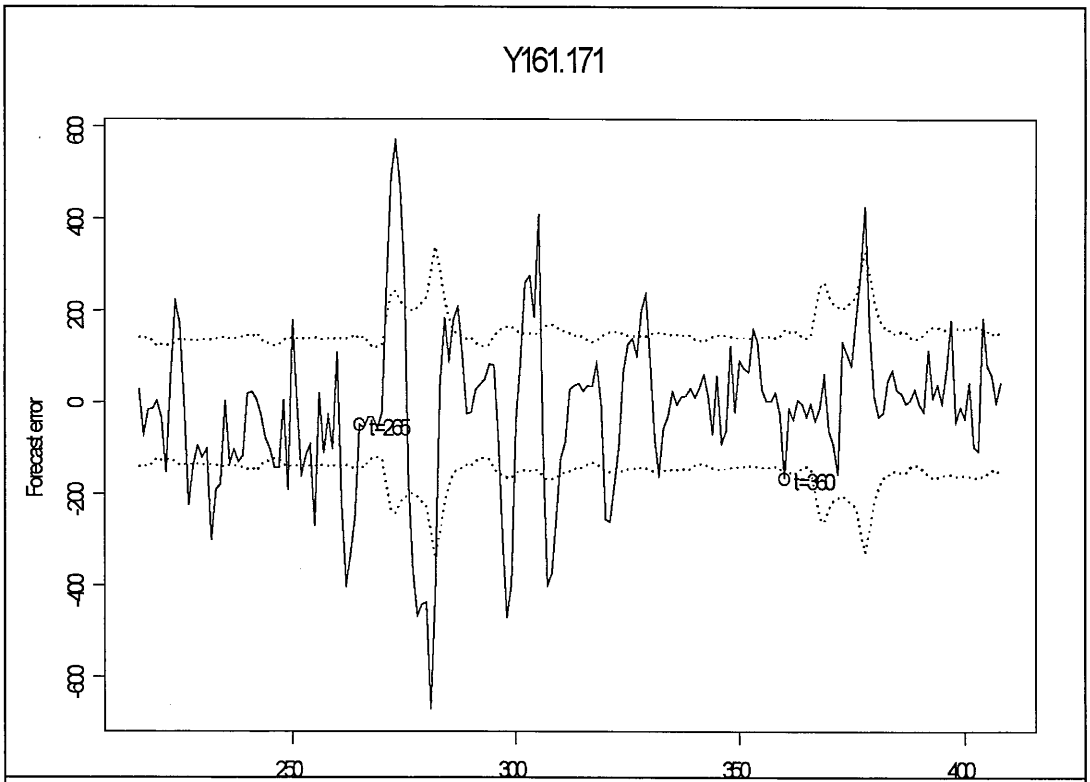

Figure 7.5.1.11 – Section of forecast errors for node Y161+Y171 after full intervention

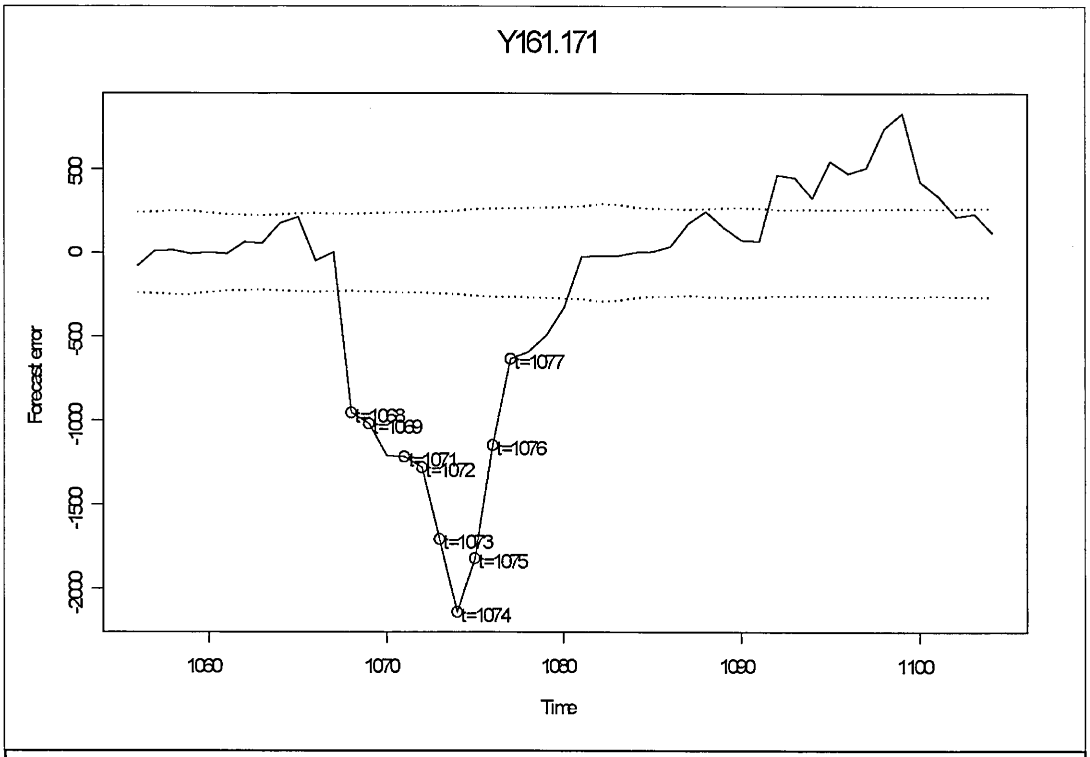

Figure 7.5.1.12 – Section of forecast errors for node Y161+171 after intervention at node Y167

in this case. For the second, lower, step, the modeller intervenes as for a short-term transient event, although the event is 'signalled' earlier.

Time	Type of intervention	Intervention value	Intervention variance
1073	Add parameter for one time point	-2000	10000
1074	Add parameter for one time point	-2000	10000
1075	Add parameter for one time point	-2000	10000

Table 7.5.1.7 – Third intervention performed at node Y(161)+Y(171)

The impact of this intervention is shown in figure 7.5.1.13. The points where intervention took place show a marked improvement in forecast variance. The surrounding nodes are as before- and may be improved through intervention but were not as unusual as the central points.

Tier 6 - nodes Y(161), Y(171)

The only intervention needed at node Y(161) is to undo the intervention at Y(161)+Y(171). As the first two events were caused by a reduced flow part node Y(171), the intervention also erroneously reduces the flow through node Y(161). Correcting this is a simple matter of reversing the intervention. It is summarised below.

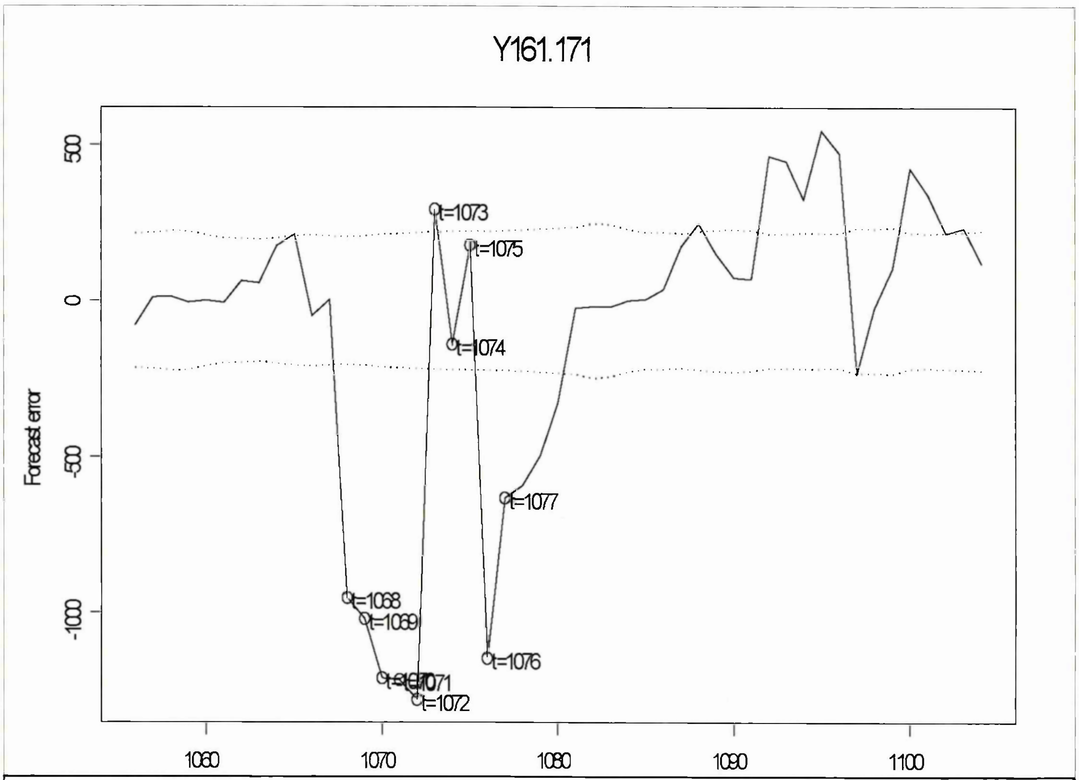

Figure 7.5.1.13 – Section of forecast errors for node Y161+171 after full intervention

Time	Type of intervention	Intervention value	Intervention variance
265	Scale ε_t by a value	×20/17	+0.005
360	Scale ε_t by a value	×17/20	+0.005
433	Scale ε_t by a value	×20/17	+0.005
576	Scale ε_t by a value	×17/20	+0.005

Table 7.5.1.8 – Intervention performed at node Y(161)

For nodes Y(161) and Y(171) there are further points where intervention might be necessary, but they do not stand out from the normal areas as markedly as areas for other intervention. A side effect of the hierarchical structure is that intervention towards the top of the network is more effective than intervention towards to bottom. Effort expended towards the top of the network affects many nodes, and thus it may be worthwhile intervening more strenuously at the top of the network. It is worth noting that most intervention in this MDM network is carried out where traffic is entering the network and little needs to be done at nodes that only receive flows from elsewhere in the network.

Tier 7 - nodes Y(162), Y(172)

Node Y(162) is another simple DLM node. There is only one time point ripe for intervention, which is a single outlier not shown in a figure.

Time	Type of intervention	Intervention value	Intervention variance
657	Disregard data point		

Table 7.5.1.9 – Intervention performed at node Y(162)

Node Y(172) is the final simple DLM node. The first event begins at time 34 in figure 7.5.1.14. This is a period of depressed flow. The later time periods can be intervened for by applying a constant intervention parameter through the period.

Time	Type of intervention	Intervention value	Intervention variance
34	Disregard data point		
35	Add parameter for one time point	-450	10000
36	Add parameter for one time point	-450	10000
37	Add parameter for one time point	-450	10000
38	Add parameter for one time point	-450	10000

Table 7.5.1.10 – First intervention performed at node Y(172)

The forecast errors after intervention are shown in figure 7.5.1.15. There is a significant improvement in the later nodes of the event.

The second event is a period of inflated flow at time 785 in figure 7.5.1.16. Unlike the previous event, there is not a consistent value the points appear to be displaced by, so all the time points are treated as outliers. The intervention at this node is summarised as:

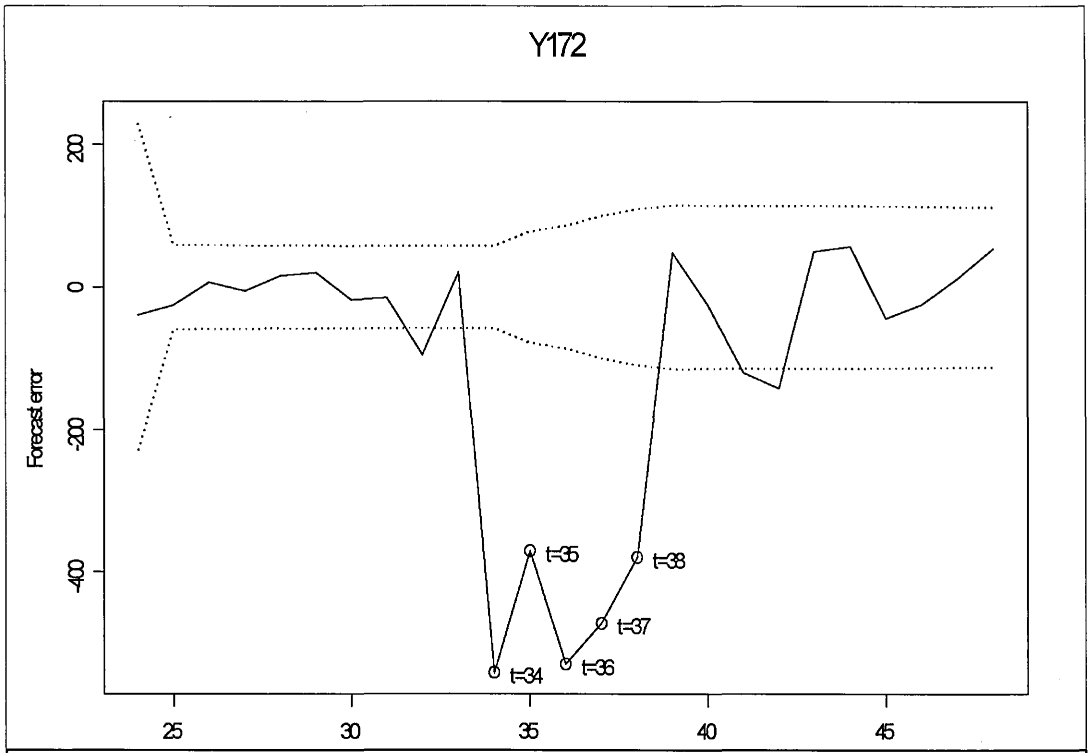

Figure 7.5.1.14 – Section of forecast errors for node Y172

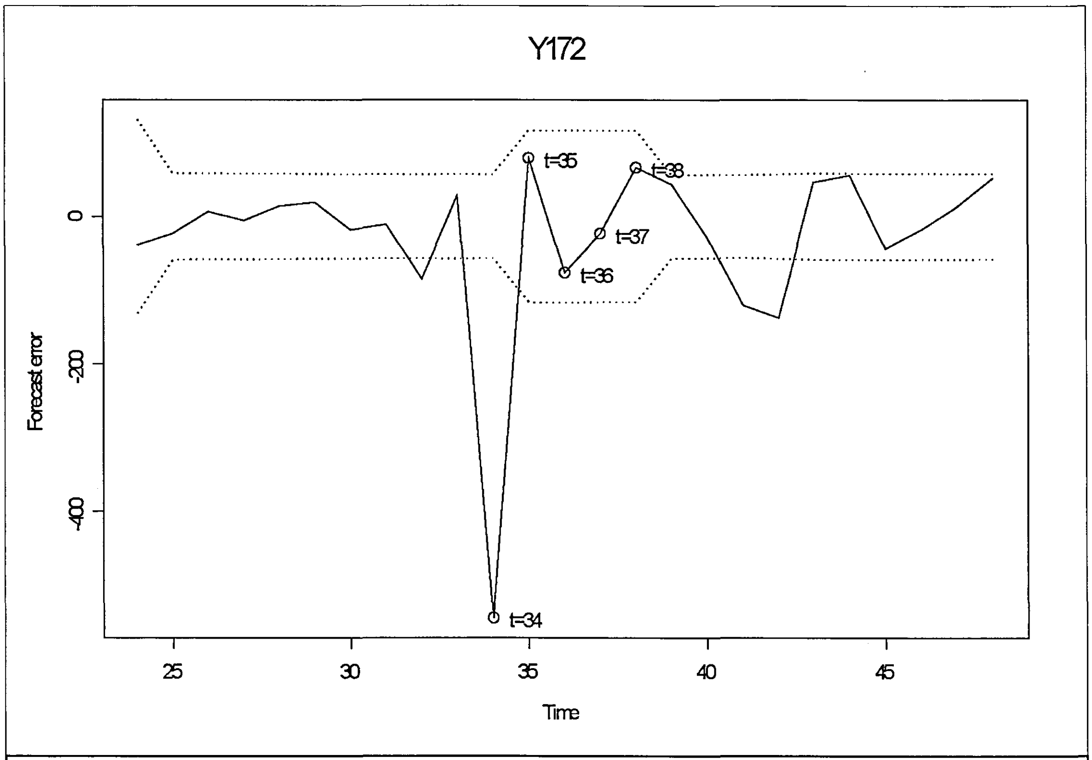

Figure 7.5.1.15 – Section of forecast errors for node Y172 after intervention

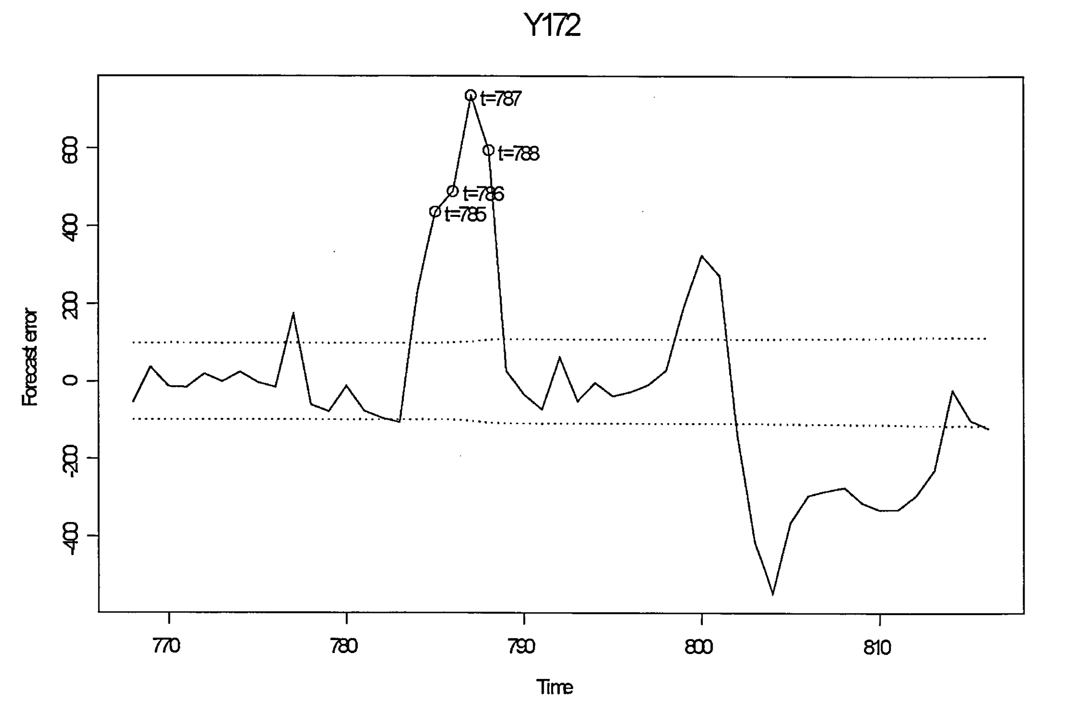

Figure 7.5.1.16 – Section of forecast errors for node Y172

Time	Type of intervention	Intervention value	Intervention variance
785	Disregard data point		
786	Disregard data point		
787	Disregard data point		
788	Disregard data point		

Table 7.5.1.11 – Second intervention performed at node Y(172)

The forecast errors after intervention are shown in figure 7.5.1.17. As with the intervention at node Y(168), although treating points during the unusual event as outliers does not improve the forecasts for them, it does improve the forecasts one day later from time 810 onwards.

Tiers 8, 9 and 10 – nodes Y(162)+Y(172), Y(163)+Y(164b), Y(163), Y(164b)

The remaining MDM nodes have no remarkable events left to intervene for.

7.5.2 Model performance with intervention

The summary statistics for the model with intervention are given in table 7.5.2.1, together with those without intervention for comparison purposes.

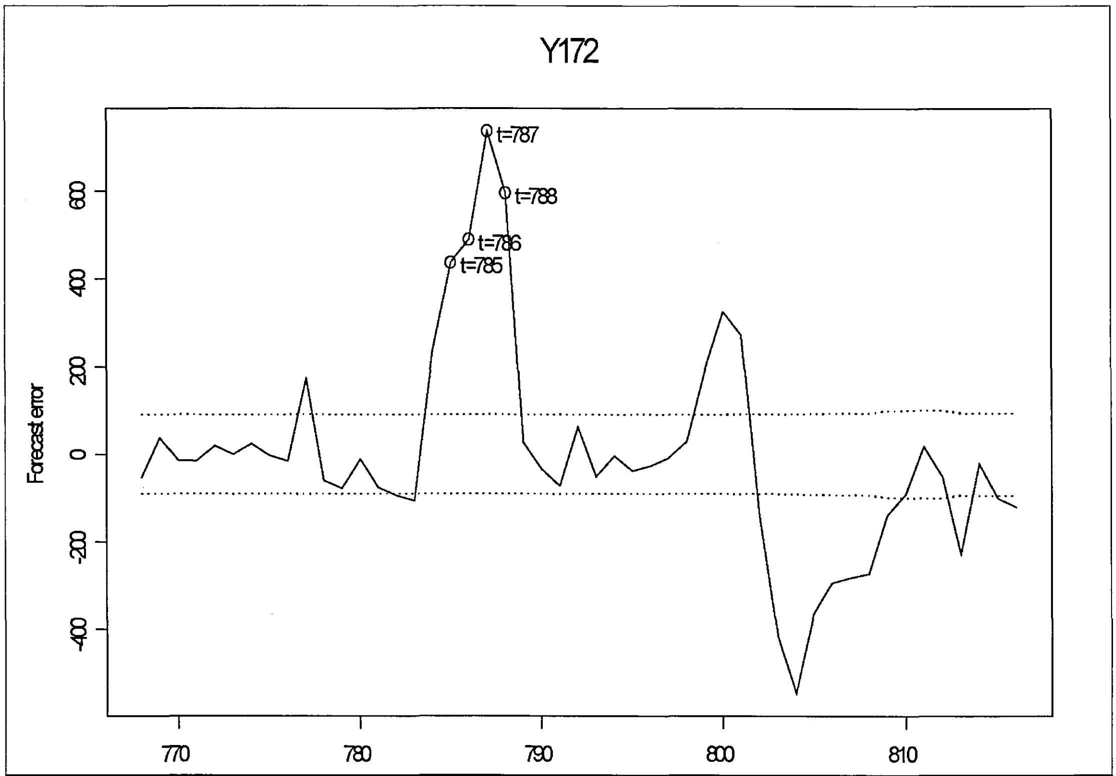

Figure 7.5.1.17 – Section of forecast errors for node Y172 after intervention

Node	Intervention		No interv.	
	MSE	MFV	MSE	MFV
162	4288	3661	4374	4190
167	36049	29896	41948	37989
169	678	641	678	641
172	5837	4035	7387	7225
161	13614	17599	11299	24662
163	982	997	937	1126
164b	5556	3158	6244	4044
168	825	771	902	1645
170a	14798	9308	17122	10810
170b	4749	3899	5425	4327
171	32095	19767	43967	26819

Table 7.5.2.1 – MDM performance with and without intervention.

The first block of nodes (162, 167, 169, 172) are simple DLMs. The intervention gains here, both in terms of forecast errors and forecast variance, are marked where intervention took place. For nodes where no compromises had to be made in order to construct the DAG- nodes 168, 170a and 170b- the gain was just as noticeable. Improvement was patchier in the remaining MDM nodes- probably as a result of the compromises in the DAG for those nodes.

Chapter 8 – Alternative MDM model

In this section, an alternative to the MDM model used previously is proposed and evaluated. It uses informal methods of modelling seasonal variance that is currently not allowed for in the model. Variance law approaches exist for DLM models (West and Harrison 1999) but they rely on relating the observation variance to the system variance or some known set of weights for observations. These are not intuitively appropriate for this sort of model as there is no natural value to use for the weights so an ad-hoc, alternative method of introducing seasonal variance is considered.

8.1 Seasonal variance

Consider equation 6.2.2:

$$Q_t(i)^* = trace\left[B_t(i)^T\left(Var(F_t(i)) + F_t(i)^* F_t(i)^{*T}\right)B_t(i)\right] + a_t(i)^T Var(F_t(i))a_t(i) + S_{t-1}(i)$$

If any parent of this node has a seasonal pattern then $F_t(i)^*$ will also have a seasonal pattern over time. Further, if this node, $Y_t(i)$, has a seasonal pattern then $a_t(i)$ will have a seasonal pattern. This means that in either of those cases the marginal forecast variance will exhibit a seasonal pattern. This pattern can be seen by examining the one-step ahead forecast variance of an MDM node, for example figure 8.1.1.

This would be expected as count data usually follow a Poisson distribution (and the variance of a Poisson distribution is equal to the mean) - so this seasonal variance is therefore not simply an artefact of the MDM model. The regression DLM nodes in the MDM have some seasonality in the forecast variance built in as shown above. This may be sufficient to encapsulate the seasonality in the forecast variance.

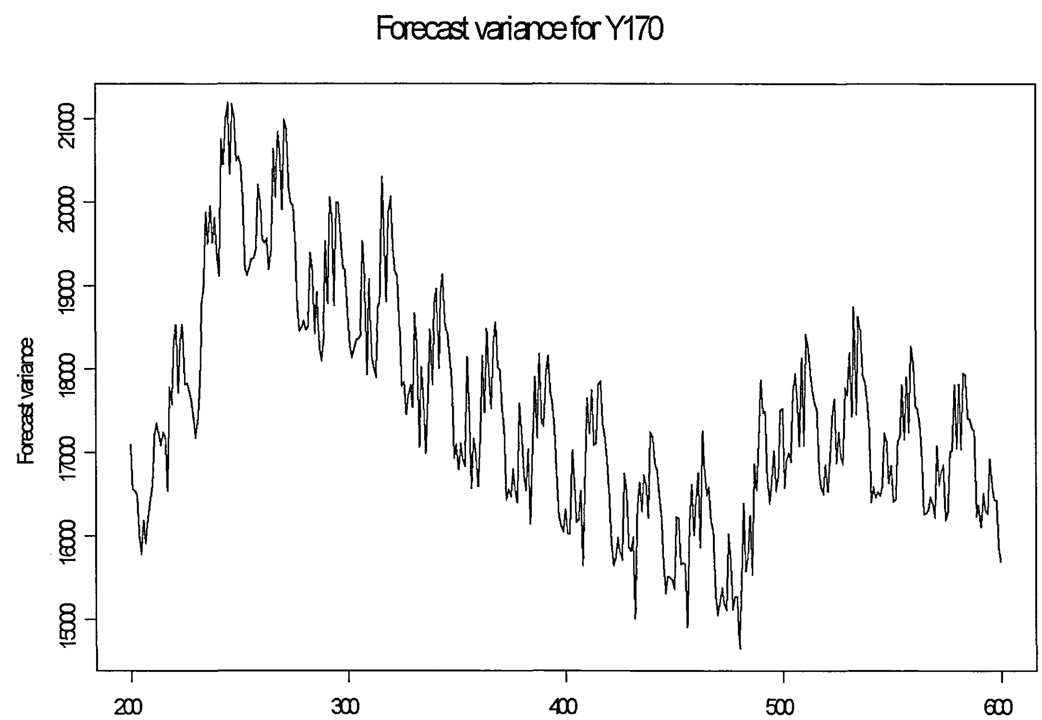

Figure 8.1.1 – Section of one-step ahead forecast variances for node Y170

However, the variance of the observation noise S_t is not made seasonal through this approach and this may be significant. In addition, the standard DLM model does not include provision for any such seasonality in the observation noise. Modelling seasonality in the forecast variance for both MDMs and DLMs may result in performance gains.

8.2 Seasonal variance estimation

Transformations of the data can remove seasonal variance in the observation noise but identifying a suitable transformation is not necessarily simple. Even when a suitable transformation can be identified it may make patterns that are clear on the original scale difficult to detect and it may inhibit interpretation of the model. This undermines the goal of simplicity and interpretability of the model parameters.

For this work, a simpler model of this seasonality is used. Specifically, a vector for the variance of the observation noise with one entry for each hour is introduced. Algebraically:

$$Y_t(i)=F_t(i)^T\theta_t(i)+v_t(i)$$
$$v_t(i)\sim N\left[0;V_{t\,mod 24}(i)\right]$$

with mod as the modulo function. This can be modelled with ease within the existing variance estimation for DLMs- simply by using 24 associated pairs of S and n parameters and updating as usual. Algebraically:

$$Q_t(i)=F_t^T(i)R_t(i)F_t(i)+S[t\,mod\,24]_t(i)$$

A side effect of this approach is that the degrees of freedom for the estimate S (given by n) will increase much more slowly than normal. This technique is now employed throughout the entire MDM model. The regression DLMs in the MDM model would have less need of this approach (as they already have seasonality in the forecast variance through different means), but they have it applied anyway for comparison purposes. Comparison of the performance of the DLM nodes with and without this technique is unaffected by the use of this technique further down the model. The form of the DLM and MDM nodes in this model is identical to that used in chapter 6, except concerning the V term.

8.3 Model performance

In the same format as before, the performance of the model is compared to the basic MDM model without intervention.

Node	MDM seasonal var		MDM no interv.	
	MSE	MFV	MSE	MFV
162	4652	3258	4374	4190
167	44454	20319	41948	37989
169	712	633	678	641
172	8690	4501	7387	7225
161	12957	22264	11299	24662
163	1433	1870	937	1126
164b	6761	6034	6244	4044
168	1006	1805	902	1645
170a	18013	9171	17122	10810
170b	6214	4559	5425	4327
171	47929	23990	43967	26819

Table 8.3.1 – Alternative and standard MDM model performances.

It can be seen that this approach is markedly worse than the standard MDM model for mean squared error. The median forecast variance is generally lower for the seasonal variance MDM. In this model the one-step ahead forecast variance varies considerably from one time period to the next. The regular MDM model essentially smoothes this pattern- leading to different median forecast variances.

The uniformly worse performance is an indication that the alternative seasonal variance model is not effective at modelling for DLM nodes or MDM nodes. This implies that more formal methods may be required to model this seasonal variance, or that the standard MDM model can model this seasonality systematically without needing any such measures in places other than at root nodes. The method also leads to less degrees of freedom in the T-distributions, which may be a factor as it will increase the variance of a node and thus make the parameters 'jumpy' in response to observations higher or lower than usual. The structure of the MDM model ensures that seasonality in the level of a non-root node and its parents creates seasonality in the variance of the node without additional work, so a more complicated solution might only need to be applied to the root nodes in the network.

Chapter 9 – Independent DLMs

The performance of the MDM model can be compared to the performance of a set of independent DLMs. DLMs allow easy intervention in the same way as MDMs do. However, a set of independent DLMs does not provide forecast covariances between the variables, limiting the information about the network that can be acquired. The lack of a relationship between the points is of signal importance when intervention is performed. Without the hierarchical structure each event may require multiple instances of intervention where an MDM, for example, might only need one. The additional information contained in a multivariate structure could itself be useful when determining how to intervene. In particular if there are two events occurring simultaneously then the MDM may indicate the effect each has distinctly. For example, if intervention at a parent node does not correct large errors for a child node it is an indication that there is a second event occurring, and its impact can be gauged from these errors. In a set of DLMs it would not be possible to say how much each event affected the child node, or even to say there were two events. The model fitted is a univariate DLM of the following form:

$$Y_t(i) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^T \theta_t(i) + v_t(i) \qquad\qquad v_t(i) \sim N\left[0; V(i)\right]$$

$$\theta_t(i) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} \theta_{t-1}(i) + w_t(i) \qquad w_t(i) \sim T_{n_{t-1}(i)}\left[0; W_t(i)\right]$$

with discounting for $W_t(i)$ and variance learning for V(i) as described in section 4.5.

9.1 Choice of priors

As in the MDM model, each DLM has priors generated from a week of data prior to the main run. The estimates of the parameters ($m_t(i)$) and the estimate of the observation noise ($S_t(i)$) are generated in this way. The degrees of freedom for the estimate of the observation noise ($n_t(i)$) are the same for each DLM, set by the length of the training data used. The variances of the estimates of the parameters ($C_t(i)$) are set to be of the form $c \times I$, where c is a large scalar chosen to ensure the model adapts quickly. These DLMs are identical in form to the root nodes in the MDM model, including seasonality. The discount factor was kept the same as for the MDM model.

9.2 Intervention

Intervention can proceed in the same manner as for the MDM model, with the previously noted exception that for a single event intervention must take place for every node it affects, instead of only at the highest tier the event occurs. In practice, the gain through intervention will be near-identical to that in the MDM model. However, the DLM model requires more quantity decisions to be made for intervention. The only locations where a intervention would be substantively different are where the MDM model has parameter reduction in the DAG. In the case of nodes 161 and 171, where a certain type of event meant intervention had to be performed twice instead of once in the MDM model (Tier 5 in section 7.5.1), independent DLMs would only require one intervention each. This still results in the same number of interventions taking place- but the MDM missed an opportunity to use intervention by tiers to reduce it. Apart from demonstrating how much extra work is required to intervene in the DLM models, there is little purpose in performing the intervention as the performance gain should be comparable.

9.3 Results

The results from a set of independent DLMs are reproduced in table 9.3.1, together with the results from the MDM model with and without intervention for comparison purposes.

Node	Indep. DLMs		MDM no interv.		MDM with interv.	
	MSE	MFV	MSE	MFV	MSE	MFV
162	4374	4190	4374	4190	4288	3661
167	41948	37989	41948	37989	36049	29896
169	678	641	678	641	678	641
172	7387	7225	7387	7225	5837	4035
161	10795	8589	11299	24662	13614	17599
163	925	854	937	1126	982	997
164b	6243	6075	6244	4044	5556	3158
168	913	965	902	1645	825	771
170a	17059	15637	17122	10810	14798	9308
170b	5439	4587	5425	4327	4749	3899
171	45410	38123	43967	26819	32095	19767

Table 9.3.1 – Performance of independent DLMs, the MDM, and the MDM with intervention.

The first half of the table is composed of nodes in the MDM model that are simple DLMs. For these nodes the DLMs and the MDM model are identical. For nodes 168, 170a and 170b the MDM model without intervention performs as well as the independent DLMs. However, the MDM model has the advantage of an innately more productive means of intervention and the capacity to provide covariances between nodes. The remainder of the nodes are nodes where a compromise had to be made drawing the DAG, but even in these nodes the performance of the two models is close. Despite performance being similar between the two models, it is worth noting that even in this relatively small network intervention in the MDM is far easier than in the independent DLMs. In a network with more tiers, or even in this network had there not been missing series, the reduction in the amount of intervention that need be performed would be even more substantial.

Chapter 10 – Independent ARIMA Models

Comparing the MDM to independent DLMs shows how it performs compared to a simpler Bayesian methodology. It is also of interest to see how it compares to a simple non-Bayesian methodology. This chapter compares the performance of the MDM with the ARIMA model- a standard non-Bayesian time series model.

10.1 Brief definition of ARIMA

ARIMA models (Box and Jenkins 1976) are commonly used time series models. An ARIMA (p,d,q) process is a stationary model defined by the hyper-parameters p, d and q. For a stationary time series X:

$$x_t=\mu+\phi_1 x_{t-1}+\phi_2 x_{t-2}+\ldots+\phi_p x_{t-p}-\theta_1\epsilon_{t-1}-\theta_2\epsilon_{t-2}-\ldots-\theta_q\epsilon_{t-q}+\epsilon_t$$

where ϵ_t is the error term at time t, and μ , ϕ_t and θ_t are parameters. Where the time series is not stationary, the differences of successive terms are modelled as a time series:

$$x'_t=x_t-x_{t-1}$$

If this still does not yield a stationary series, then differences can be taken again. The number of times the difference is taken is given by the hyper-parameter d. The parameters are then estimated using least squares.

ARIMA models can also deal with seasonality, by introducing an additional ARIMA component that operates at a specific lag. The notation is then ARIMA (p,d,q)×(p',d',q')$_r$, where r is the lag required. The equation is then:

$$\begin{aligned}x_t=&\mu+\phi_1 x_{t-1}+\ldots+\phi_p x_{t-p}-\theta_1\epsilon_{t-1}-\ldots-\theta_q\epsilon_{t-q}\\&+\phi'_1 x_{t-r}+\ldots+\phi'_{p'}x_{t-rp'}-\theta'_1\epsilon_{t-r}-\ldots-\theta'_{q'}\epsilon'_{t-rq'}+\epsilon_t\end{aligned}$$

where ϕ' and θ' are the parameters associated with the seasonal part of the model.

10.1.1 Ramifications of ARIMA in the network

The predictor of an ARIMA (p,d,q) process is the limiting form of a certain class of DLMs where no intervention takes place (West and Harrison 1999). However, these limiting forms are not reached in applications where intervention takes place. There are techniques for performing intervention in ARIMA models, see for example Mélard and Pasteels (2000) or Bianchi, Jarrett and Hanumara (1998). However, the intervention is exogenous to the ARIMA model and is not incorporated in the same way that intervention in a DLM model can be. The Bayesian methodology of a DLM-based approach allows smoother integration of the intervention and provides priors for the intervention parameters. Intervention is highly dependent on the information available to the forecaster, so while performing intervention in the ARIMA model would indicate the kind of forecasting improvements that could be made, it would not be a like-for-like comparison with intervention in the MDM model.

ARIMA models also assume constant variance. This assumption falls down for this application in two ways- the structure of the system introduces a seasonal variance (as can be seen in equation 6.2.2) and examination of the data shows that even the variance due to the 'noise' parameter seems to follow a seasonal pattern (as discussed in chapter 8). Consequently, the standard ARIMA model is not best suited to this problem.

A more complex approach for this application may be based on ARIMA methodology in the same way as the linear MDM is based on DLMs. For simplicity here, however, only a standard ARIMA approach is employed for comparison purposes, as the more in-depth approaches are beyond the scope of this thesis. Standard ARIMA, by contrast, is well supported in statistical computer packages and can serve as a simple non-Bayesian comparator.

In common with the independent DLM models no use is made of the links between the time series. ARIMA models also assume the model parameters are constant over time, which is not guaranteed in this application.

10.2 Model selection

The autocorrelation and partial autocorrelation for a typical node are shown in figure 10.2.1. There is not an obviously correct model, although an ARIMA (0,0,2) model might be appropriate before seasonality is considered as there are two large initial lags in the partial autocorrelation plot and the autocorrelation plot appears to be decaying slowly. The correlation at many lags is statistically significant (as indicated by the horizontal bars), but this is most likely because the data are not well suited to this model.

In order to establish the most suitable model, a more formal method can be used. Comparing models through the Akaike Information Criterion suggests that $(1,0,0)\times(1,0,0)_{24}$ is the most suitable seasonal ARIMA model. Note that the seasonal component is at lag 24- the same lag that the MDM and DLM models use- but this model also takes into consideration the lag 1 term.

This leads to a model of the form:

$$x_t=\mu+\phi_1 x_{t-1}+\phi'_1 x_{t-24}+\epsilon_t$$

The parameters for lag 1 and lag 24 were estimated using a standard function in the language S-Plus that uses a convergence-based method.

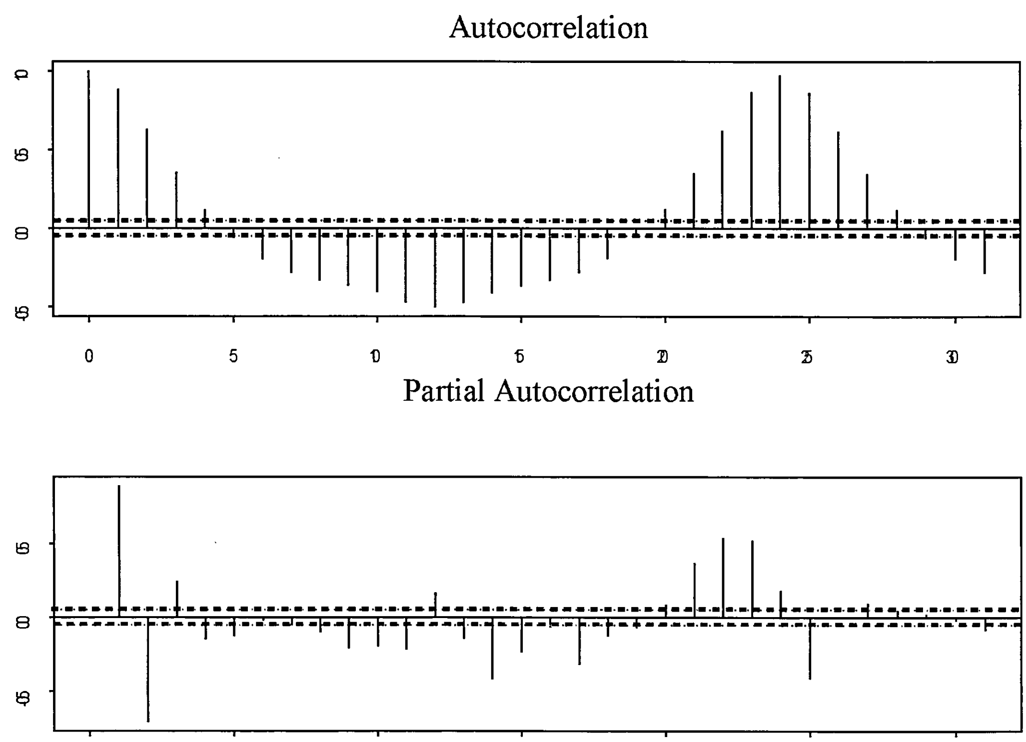

Figure 10.2.1 – Autocorrelation plots for Y167

10.3 Results

Once the model is established, generating results is a simple matter of calculation. The parameters converged to in the model were $\phi_1 \approx 0.5$ and $\phi'_1 \approx 1$. This implies that the lag 24 term is most significant, but the lag 1 term does influence the observed value (as one would expect). The results obtained, as compared to the MDM model without intervention are given in table 10.3.1.

Node	**ARIMA**		**MDM no interv.**	
	MSE	**MFV**	**MSE**	**MFV**
162	5215	5127	4374	4190
167	39408	38548	41948	37989
169	918	907	678	641
172	6183	6167	7387	7225
161	9137	8894	11299	24662
163	1314	1314	937	1126
164b	4851	4813	6244	4044
168	1064	1046	902	1645
170a	16025	15721	17122	10810
170b	5557	5440	5425	4327
171	23682	23086	43967	26819

Table 10.3.1 – Performance of the independent ARIMA models and the MDM without intervention

The parts of the MDM model that are simple DLM models (the upper part of the table) perform about as well as the independent ARIMA models. For most of the MDM nodes, performance is about the same as for independent ARIMA models. Differences between them could be due to the suitability of the ARIMA technique for each particular node. Some nodes may be more constant over time than others, or may have lag 1 values of greater impact on their level. The one node for which performance in the MDM was markedly worse than independent ARIMA models was node Y(171). This node was in part of the network where compromises had to be made to the DAG, and was also a deterministic twin so was not modelled directly. This poor performance may be corrected by making it the modelled node and Y(161) the deterministic twin.

It is interesting to note that the ARIMA models use more information than the DLM-based models- specifically the data at the previous time point- but offer no general performance gain except for one node that is known to have problems. The DLM and MDM models have the natural advantage of the greater ease of expert intervention. Unlike the independent DLM models, the ARIMA models are not restricted to steady seasonals thus a direct comparison may not be made. More importantly, the ARIMA estimation uses the entire data set to estimate the parameters and then finds the forecast errors over that same set. The DLM-based models only use the data up to that time for forecast the next data point. However, this may not be an advantage if the model is not stationary (after accounting for seasonality). Thus the comparison is not entirely fair, although it is possible to update the forecast function of an ARIMA model in a similar way to that of DLMs (Butler, 1999). However, even though the ARIMA model uses the entire data set the MDM model is competitive for nodes where no compromises were made formulating the DAG.

Chapter 11 – Discussion and Further Research

This chapter has three sections. Firstly there is an analysis of all the methods presented for forecasting traffic flow and a discussion of how they compare. The second section evaluates the MDM methodology presented here and outlines its advantages and problems. The last summarises areas for future research around the problem and the model. Whitlock (1999) applied an MDM model to the same data set, however, a different DAG was used that did not take into account deterministic twins. Also, MCMC was used to simulate missing data instead of amending the DAG. MCMC was required because the DAG employed by Whitlock did not allow for MDM nodes that had flows from both inside and outside the network. This avoids the problem of covariance between entry flows, but represents a more restrictive subset of MDM models and requires MCMC in order to deal with missing data. Both approaches would be appropriate in different circumstances. Whitlock introduced artificial events into the data to demonstrate intervention, but in this thesis the existing unusual events (although devoid of context) are used for this purpose.

Cassidy, Anani et al. (2002) provide evidence that congestion on one branch of a network causes lowered flow of a peer branch- so traffic congestion causes higher correlation between siblings in the network. Maunch and Cassidy (2002) suggest this oscillation travels upstream from the congested area. In the MDM, the data would first signal this event upstream of the congestion, and if intervention is performed reactively then intervention at that point might be the only intervention needed to match both these behaviours. The overall reduced flow means that the two siblings also have reduced flow. However, if the intervention is to be performed actively, then the forecaster must ascertain the first earliest point at which to intervene. This might be non-trivial, however, the structure of the MDM ensures that the implementation of the intervention remains simple.

11.1 Comparative performance

The performance of the basic MDM against other methodologies is given in table 11.1.1. The nodes are divided into three types, root nodes (in the first section), nodes where the MDM broke proper causality relationships (the second section) and nodes where it did not (the final section). The method with the best mean squared error for a node and any methods with a mean squared error within 5% of the best have been shown in boldface.

Node	MDM, no intervention		Independent DLMs, no interv.		ARIMA	
	MSE	MFV	MSE	MFV	MSE	MFV
162	**4374**	4190	**4374**	4190	5215	5127
167	41948	37989	41948	37989	**39408**	38548
169	**678**	641	**678**	641	918	907
172	7387	7225	7387	7225	**6183**	6167
161	11299	24662	10795	8589	**9137**	8894
163	**937**	1126	**925**	854	1314	1314
164b	6244	4044	6243	6075	**4851**	4813
171	43967	26819	45410	38123	**23682**	23086
168	**902**	1645	**913**	965	1064	1046
170a	17122	10810	17059	15637	**16025**	15721
170b	**5425**	4327	**5439**	4587	**5557**	5440

Table 11.1.1 – Performance of the MDM without intervention and independent DLM and ARIMA models.

It can be seen that for root nodes, where the MDM is equivalent to the independent DLM models, performance is broadly similar between those and ARIMA methodology.

In the second section, MDM and DLM methodologies performed similarly, but for most of the nodes less well than ARIMA methodology. In particular they performed badly for node 171. This is probably due to the unusual activity in node 171, identified in section 7.5.1. The activity is a period of depressed flow, with clear start and end points. The ARIMA method will suffer poor performance around these end points, but as the daily pattern remains the same will perform well during these periods. The MDM and DLM, however, will slowly adapt to the unusual activity then slowly adapt back when it ends leading to poor performance over a longer period. Intervention should prevent or alleviate this problem. Even with the loss of causal relationships, the MDM performs as well as independent DLMs for these nodes. This suggests that the MDM is only losing information from the relationships between the sites and not breaking down completely. It also suggests that when such compromises may have to be made when formulating the DAG, it is feasible to model such nodes with independent DLMs, then pick up the MDM structure from then on in. This eliminates any problems in the formulation of the DAG for an MDM network, but loses the benefits of intervention by tiers. The forecaster does, at least, have this option. In this case, such points are at the end of the flow structure so it would be equivalent to modelling all the nodes in the second section with their independent DLM equivalents. The competitive performance of the ARIMA model for these nodes suggests that the lag 1 term the ARIMA includes may be significant enough to warrant inclusion in a future MDM model.

In the third section, performance of all three models is again broadly equivalent.

The median forecast variance is similar to the mean squared error for most models and nodes, as we would expect as this is count data and thus Poisson. The Normal distribution should be a good approximation of this as the observed counts are high, as is borne out by the results. The exception is in the second category, where the MDM model produces median forecast variances vastly different from the mean squared errors. This would appear to be an equalising effect between the siblings in this part of the model. Nodes 161 and 171 are siblings, with 161 having a high median forecast variance and 171 having a low median forecast variance. Nodes 163 and 164b show a similar pattern. The precise nature of the smoothing effect between these pairs does not seem to be a simple linear trade-off. Its presence can be explained by examining the DAG used to predict them (figures 7.1.2 and 7.2.2). In each case the pair is an MDM node and its deterministic twin with their sum as their parent. It could be that the MDM model does not adequately determine how much of the uncertainty regarding the sum should be passed to each of the children. This is another drawback of breaking the causality relationships when formulating the model. A similar situation exists with nodes 170a and 170b, which in the DAG both stem from a parent which is their sum. In this case, however, there is no break in causality and the effect appears to be a reduction in forecast variance for both rather than a trade-off. The production of forecast bounds narrower than those given by a Poisson approach is curious. This could be a result of the MDM exploiting the causal relationships beneficially or an artefact of the seasonal variance the MDM produces. Further exploration of the effect may be of interest.

The above table only examines how the MDM and DLM models perform without intervention. It does establish that even without intervention the MDM is a competitive model against others of similar complexity. Once intervention is performed, the MDM compares to ARIMA methodology as shown in table 11.1.2.

Node	MDM interv.		ARIMA	
	MSE	MFV	MSE	MFV
162	**4288**	3661	5215	5127
167	**36049**	29896	39408	38548
169	**678**	641	918	907
172	**5837**	4035	6183	6167
161	13614	17599	**9137**	8894
163	**982**	997	1314	1314
164b	5556	3158	**4851**	4813
171	32095	19767	**23682**	23086
168	**825**	771	1064	1046
170a	**14798**	9308	16025	15721
170b	**4749**	3899	5557	5440

Table 11.1.2 – Performance of the MDM with intervention and independent ARIMA models.

The performance of the MDM in the first category of nodes has markedly improved after intervention. In particular node 167 has a 14% reduction in mean squared error after intervention. Node 167 has the heaviest intervention applied to it and it is a good indication that intervention was performed satisfactorily. In particular, the MDM now outperforms the ARIMA model in all nodes in this category. Of course, these nodes are simply independent DLMs so that model would show a similar improvement.

In the second category of nodes, the MDM still performs less well than ARIMA methodology. However, the mean squared errors have improved. The median forecast variances are still dissimilar to the mean squared errors except now for node 163.

The third section shows the MDM methodology as a clear winner. Once again the median forecast variances are lower than would be suggested by a Poisson distribution.

This is not a complete like-for-like comparison as no intervention has been performed for the ARIMA model. However, ARIMA methodology does not have a simple mechanism for intervention. Major improvements could only be gained by extensive revision of the methodology. The improvements in the performance of the MDM were gained by simple ad hoc intervention. There may be further improvements in performance with more rigorous intervention. However, as with all cases of expert intervention the capability of the expert is important. Without context to place the data in or a trial of the methodology in real-time it is difficult to gauge how well intervention could be performed. However, the capacity for intervention is the key advantage of the Bayesian approach to this application. In addition, the hierarchical structure of the MDM and the intervention by tiers approach offers advantages over intervention in a series of independent DLMs.

The adapted MDM model with seasonal observation variances from chapter 8 is compared to the standard MDM and independent ARIMA models in table 11.1.2.

Node	Adapted MDM		MDM no interv.		ARIMA	
	MSE	MFV	MSE	MFV	MSE	MFV
162	4652	3258	**4374**	4190	5215	5127
167	44454	20319	41948	37989	**39408**	38548
169	712	633	**678**	641	918	907
172	8690	4501	7387	7225	**6183**	6167
161	12957	22264	11299	24662	**9137**	8894
163	1433	1870	**937**	1126	1314	1314
164b	6761	6034	6244	4044	**4851**	4813
171	47929	23990	43967	26819	**23682**	23086
168	1006	1805	**902**	1645	1064	1046
170a	18013	9171	17122	10810	**16025**	15721
170b	6214	4559	**5425**	4327	**5557**	5440

Table 11.1.3 – Performance of the adapted and original MDMs without intervention and independent ARIMA models.

The performance of this model is uniformly worse than either the original MDM or the independent ARIMA models. This may be due to the decreased precision in the estimation of the observation noise variances in the adapted model. It is clear that if seasonal variances are to be accounted for in either DLMs (for the root nodes) or the MDM in general a different approach is needed. However, as an MDM model includes terms that will naturally produce seasonality in the forecast variances it may not be necessary to apply them to MDM nodes. The DLMs used for root nodes might still benefit from such an approach.

11.2 Evaluation of the methodology

The MDM model performed adequately on the basis of mean squared errors, but there are other aspects to consider when assessing its effectiveness.

An important issue concerning model validation is the choice of distributions. For root nodes and other inflows into the network, the natural distribution to choose is the Poisson. A Normal approximation (and, by extension, the T distribution used with variance estimation) is appropriate as long as the mean of the distribution is relatively high. However, during the night some nodes have very low observed values which might bring into question the validity of a Normal approximation.

Experimentation with formulation of the DAG was performed to account for unknown covariances between entry points (see section 7.1). This experimentation showed that when the model fails in this circumstance it fails in a way indicative of overparameterisation, with unbounded growth in the forecast variance. When this happened, it was during the parts of the cycle where flows were very small that showed the highest variance growth. It could be possible that these low values are the cause of this seeming overparameterisation instead of the model formulation. Forecasts for these time points might be considered of little importance by the forecaster, and the DLM and MDM could operate with these time points omitted. Future work could examine how this changes the behaviour of the model in those parts of the DAG that are causing problems. Problems arising from the covariances between entry points could also be solved by considering the network as an aggregation of the noisy AND-OR-NOT Bayesian networks examined in Schubert (2004), or by introducing an overall traffic level component as part of a dynamic hierarchical model of the kind described in Gamerman and Migon (1993).

Similarly, the conditional distribution for an MDM node would be more naturally a Binomial distribution. However, this could only be implemented if the forecast distributions for the parent nodes had integer values. Additionally, such measures may make the marginal distribution more difficult to calculate. A Normal approximation to the Binomial is appropriate is most cases, but for some MDM nodes there are quite extreme values for p which brings into question the suitability of the approximation. However, as the forecasts for MDM nodes consist only of point forecasts and variances the symmetry of the distribution is not part of those calculations. The approximation is likely to be 'good enough' for the purpose of forecasting future traffic flows and examining the relationships between them.

The independence of parameters is important in an MDM model. The formulation of the DAG guarantees the independent updating of parameters for MDM nodes, but does not help with the parameters for root nodes. The model presented here assumes independence between inflows. This is not a credible assumption and contributed to the problems with compromised sections of the DAG. Finding the covariance between root nodes is a significant problem in its own right. Ad hoc numerical methods could provide a value for this covariance, but there is no Bayesian formulation for covariance between variables. As such obtaining some simple distribution that can easily be incorporated into the Bayesian framework of the MDM is not possible. The inverse Wishart distribution may be of use but this would have to be examined in depth in future work.

The formulation of the DAG is an important concern in MDM models, and the DAGs produced here are shown to have problems where there are inflows into the network not at root nodes. However, this situation can be avoided by design simply by placing the counting points such that no nodes have inflows from outside the network as well as parents. In this application the full network of counting points meets this criterion- it is only because some of these points have no data available for them that the problem arises in the pruned network. If there were some external means of determining the covariance between entry points, a composite MDM node with parents and an in-flow would still be tractable. In some applications this may be possible, but the only recourse in this application would be MCMC methods, which are computationally intensive and not really suitable for a real-time system. A method for determining whether a multivariate conditional Normal model can be decomposed into univariate conditional Normal models is given in Didelez and Edwards (2004). This may be of great use in formulating DAGs for an MDM and finding parts of the graph where univariate conditional models cannot be created without further refinements.

The key advantage of an MDM technique is intervention. A traffic flow pattern of this kind is likely to be noisy, and intervention may be needed frequently. The tiered structure of the MDM network and the intervention by tiers it engenders offer a powerful way of improving the performance of the model. The larger the network used, the more powerful this technique becomes. It does not reduce the amount of monitoring that needs to be done to check whether intervention is necessary, but it makes each intervention more powerful and will reduce the number of interventions (and hence the amount of elicitation) that has to be made.

A practical consideration concerning intervention is where the cause of an event takes place. Bertini and Cassidy (2002) demonstrate that bottlenecks need not form immediately in the vicinity of an entry of exit slip-road. In the context of the

MDM, this matters little as the event and the intervention for it deals only with the flow at the counting points. The position of the bottleneck will affect how quickly the congestion will percolate up stream. Mechanistic models would need to account for this, but the heuristic method the MDM model uses for intervention eliminates the need for a formal system in favour of expert opinions.

The second most important consideration in formulating this model was that of simplicity. By making the parameters understandable quantities it makes it easier for the forecaster to intervene, not just in terms of making the quantities to be added to the model simple, but also making the decisions regarding how to intervene more straight-forward. The forecaster can tell at a glance what the parameters are at a certain point at a certain time without having to perform any transformations to get real-world figures. The use of an easily used base model, the DLM, makes the implementation of the model easier and makes the mechanics of intervention simpler.

Another useful property is that by obtaining covariances between quantities of interest (see chapter 6) it is possible to form a new model when some structural change is made in the network. Traffic monitoring can also be used to improve flow by changing some aspects as the sequence of traffic signals (see Cassidy, Anani and Haigwood 2002) or temporary speed restrictions. Intervention can be used when this takes so the model adapts to the new traffic conditions. A system without intervention or some equivalent mechanism might give poor performance whenever its forecasts were used to change aspects of the network.

These qualities are of great utility where a model is a candidate for use in a real-time system. Such a model must be able to be evaluated as fast as the data are gathered. Once the initial work establishing the DAG is in place, the MDM model can do this, even including the time needed to intervene.

In particular, the avoidance of MCMC is important both for simplicity and for practicality in use as a real-time system. Previous approaches such as Whitlock and Queen (2000) and Tebaldi, West et al. (2002) used MCMC techniques as part of a Bayesian framework. Whitlock (1999) used MCMC to simulate missing data in the same network modelled here, which would not be practical in a real-time system. Tebaldi, West et al. modelled a stretch of freeway as a chain of counting points, with entries to the network found using a smoothed curve generated using MCMC. In that model, the data were minute-by-minute. MCMC is a powerful tool, but has drawbacks in this application. The reliability of an MCMC simulation is dependent on the time the simulation was run for. In a real-time system, there may not be enough time between data points to perform it satisfactorily. In addition, repeated use will greatly increase the chance that it produces an erroneous result over the course of running the model.

The MDM model is a powerful tool in this application and when applied correctly offers many advantages over other approaches of similar complexity. In particular, its many advantages in terms of intervention are attractive in any application where intervention will be an important part of modelling. Although the MDM is competitive as compared to independent DLMs or ARIMA models, it may not perform as well as more complicated approaches outlined in chapter 1. However, in the same way as ARIMA methodology forms the basis of some of these advanced approaches, the MDM model described here could form the basis of a more advanced Bayesian system.

11.3 Future work

There are many areas in which future research might prove fruitful with regards to the MDM. The most basic would be to restore the full week to the model. At the moment the model assumes continuity between Thursday and Tuesday not only for standard running of the model but for intervention. The inclusion of the full week using dummy variables as in Whitlock (1999) would be helpful. In particular, the dummy variables approach may be simpler than the approach needed for an ARIMA model to model the full week.

Given the significance of intervention in the MDM model it would be useful to have some formal monitoring method available. General techniques of formal monitoring (such as CuSums) could be used for a traffic flow application but it would be useful to have a technique specifically tailored to it. Formal monitoring would not only make intervention less ad hoc than the method used in this thesis, but could also categorise unusual events according to their behaviour. This would be of great use when determining how to intervene in response to such an event. Formal monitoring could also be used to ensure that intervention, when performed, is performed effectively. An additional approach would be to incorporate event detection methods exogenous to the MDM, using such methods as those in Coifman (2003a), Holland (1998) or Lin and Daganzo (1997).

There are two ways of assessing the effectiveness of intervention in the MDM model compared to independent DLMs. One way, as shown earlier, is to compare how many interventions need to be made for a particular period of unusual activity. This, through intervention by tiers, shows that intervention in an MDM is more effective than that in a DLM. The second method would be to compare the MSE for nodes in the models with a fixed number of interventions for each model. This would show how much better the forecasts obtained are given the same effort expended on intervention. However, this would require some means of ordering periods of unusual activity so the most significant are intervened for in preference to others. This may be possible using formal monitoring methods should they be developed.

The deterministic twins presented in chapter 6 have the limitation that they cannot be applied on situations where the directly modelled twin has more than one parent. This limitation may be removed by considering a more general approach to the constraints imposed in such situations, possibly along the lines of the constrained Kalman filter in Pandher (2002).

The seasonal pattern in the variance needs more examination. The MDM nodes account for some seasonal pattern in the variance, but the DLMs used for the root nodes do not. In particular, the Poisson characteristic that the variance equals the mean could be enforced by changing the DLM and MDM nodes as follows:

$$Q_t^* = f_t^*$$

This eliminates the need to estimate V through any means. This method should be applied to the marginal forecasts, which means calculating V_t from equation 6.2.2. Difficulties might arise if this would force V_t to take a negative (hence unacceptable) value. The formulation of such a model would be of considerable interest as it would be of an essentially simpler form than the MDM model presented here!

For both the effective use of the general DLM-based MDM model and the production of the covariance matrix for the parameters, it is necessary to find covariances between inflows. As mentioned in the previous section this is a suitable candidate for future work. One means that allows estimation of in-flows is through the introduction of data regarding vehicle reidentification. These data need not be complete in order to be useful (Coifman 2003b). Covariates in DAGs for social networks have been included in Gill and Swartz (2004) and may be be introduced in a similar fashion in this application. Another possibility is the use of nested threshold autoregressive models (Astatkie, Watts and Watt 1997)- which decompose a non-linear dynamic system into linear subsystems. This is the same principle as the MDM uses, although the NeTAR models use multiple explanatory inputs and a stateful system. However, these two properties may form a bridge between the MDM method here and the numerical methods of forecasting found in traffic literature if they are used for the entry points to an MDM model.

One missing aspect from the theory of the model is a proof of convergence for DLMs with non-constant F. Intuitively, this should be possible using a bounding argument, but at present it has not been shown. Attempting to write the observation matrix T (equation 4.2.1) for such a model leads to variables appearing in it. It is not clear how to treat them when determining whether the model is observable or not. It seems likely that the proof will impose properties on the form of an MDM node and the values in the F matrix, particularly with regards to covariance. That there are conditions in which convergence occurs does not seem to be in doubt, but without full understanding of the requirements of convergence it cannot be guaranteed. Determining these requirements may well provide insight into the problems with portions of the DAG and how to avoid them.

It can be seen that the MDM is an important extension to a general regression DLM over time. Consider the model of Tebaldi, West et al. (2002). Each node in that model was regressed on its parent node at different time lags. The lags required were established through separate analysis. In such a model, as the time periods become larger or the distances smaller, the model may have to regress at lag 1 and then need to regress at lag 0. This lag 0 regression is what the MDM model allows. A general DLM/MDM model could incorporate regression at any non-negative lag. In this application, the model could be extended to include regressing on the parents at lag 1. This would attempt to account for vehicles between counting points when the hour rolls over- a possible reason for better ARIMA forecast for some nodes. The MDM could be a useful extension to any hierarchical DLM model when used in this way. This general approach could be used with smaller sampling periods, as used in other traffic flow research. Aggregated counts can still be found and modelled as in Schmidt and Gamerman (1997).

Additionally, the ARIMA methodology suggests that alternative G-matrices may merit further study, where the parameter for time t depends not only on the parameter at time t-24, but also on time t-1. The ARIMA model fitted in chapter 9 suggests that a G-matrix of the form:

$$G'=\begin{bmatrix} 0.5 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & 0 & 1 \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

may be suitable. However, representing AR models as DLMs generally requires a transformation of the series to be zero-mean (West and Harrison 1999), and transformations may fail to meet the goal of making the model readily interpretable as established in chapter 1.

References

Ahmed, S. A. and Cook, A. R. (1979). “Analysis of freeway traffic time-series data by using Box-Jenkins techniques.” Transportation Research Record No. 733, 1-9.

Akamatsu, T. (1996). "Cyclic flows, Markov process and stochastic traffic assignment." Transportation Research Part B **30**(5): 369-386.

Astatkie, T., Watts, D. G. and Watt, W. E. "Nested threshold autoregressive (NeTAR) models." International Journal of Forecasting **13**: 105-116.

Banks, J. H. (2003). "Average time gaps in congested freeway flow." Transportation Research Part A **37**(6): 539-554.

Bertini, R. L. and Cassidy, M. J. (2002). "Some observed queue discharge features at a freeway bottleneck downstream of a merge." Transportation Research Part A **36**(8): 683-697.

Bianchi, L., Jarrett, J. and Hanumara, R. C. (1998). "Imrpving forecasting for telemarketing centers by ARIMA modelling with intervention." International Journal of Forecasting **14**(4): 497-504.

References

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. San Fransisco, Holden-Day.

Brockwell, P. J. and Davis, R. A. (1987). Time Series: Theory and Method, Springer-Verlag.

Brooks, S. P. (1998). "Markov Chain Monte Carlo and its application." The Statistician **47**(1): 69-100.

Bunker, J and Troutbeck, R. (2003). "Prediction of minor stream delays at a limited priority freeway merge." Transportation Research Part B **37**(8): 719-735.

Butler, N. A. (1999). "Updating the forecast function of ARIMA models and the link with DLMs." Journal of Forecasting **18**(4): 275-284.

Camus, R., Cantarella, G. E., et al. (1997). "Real-time estimation and prediction of origin-destination matrices per time slice." International Journal of Forecasting **13**: 13-19.

Carey, M. and Subrahmanian, E. (2000). "An approach to modelling time-varying flows on congested networks." Transportation Research Part B **34**(3): 157-183.

Cargnoni, C., Muller, P and West, M. (1997). “Bayesian Forecasting of Multinomial Time Srries Through Conditional Gaussian Dynamic Models.” Journal of the American Statistical Association **92**(438): 640-647.

Cassidy, M. J. and Maunch, M. (2001). "An observed traffic pattern in long freeway queues." Transportation Research Part A **35**(2): 143-156.

Cassidy, M. J., Anani, S. B. and Haigwood, J. M. (2002). "Study of freeway traffic near an off-ramp." Transportation Research Part A **36**: 563-572.

Coifman, B. (2003a). "Identifying the onset of congestion rapidly with existing traffic detectors." Transportation Research Part A **37**(3): 277-291.

Coifman, B. (2003b). "Estimating density and lane inflow on a freeway segment." Transportation Research Part A **37**(8): 689-701.

Consonni, G. and Leucari, V. (2001). "Model determination for directed acyclic graphs." The Statistician **50**(3): 243-256.

Cowell, R. G., Dawid, A. P., et al. (1999). Probabilistic Networks and Expert Systems. New York, Springer-Verlag.

Daganzo, C. F. (1995). "Requiem for second-order fluid approximations of traffic flow." Transportation Research Part B **29**(4): 277-286.

References

Danech-Pajouh, M. and Aron, M. (1991). "ATHENA: a method for short-term inter-urban motorway traffic forecasting." Recherche Transports Sécurité (English Issue) **6**: 11-16

Didelez, V. and Edwards, D. (2004). "Collapsibility of graphical CG-regression models." Scandinavian Journal of Statistics **31**(4): 535-551.

Dougherty, M. S. and Cobbett, M. R. (1997). "Short-term inter-urban traffic forecasts using neural networks." International Journal of Forecasting **13**: 21-31.

Eldor M. (1977). "Demand predictors for computerized freeway control systems." Proceedings of the 7th International Symposium on Transportation and Traffic Theory, Kyoto, Japan, 341-358.

Faria, A. E. and Smith, J. Q. (1997). "Conditionally Externally Bayesian Pooling Operators in Chain Graphs." The Annals of Statistics **25**(4): 1740-1761.

Fraser D. A. S. and Haq, M. S. (1969). "Structural Probability and Prediction for the Multivariate Model." Journal of the Royal Statistical Society Series B **31**(2): 317-331.

Freeland, R. K. and McCabe, B. P. M. (2004). "Forecasting discrete valued low count time series." International Journal of Forecasting **20**(3): 427-434.

Gafarian, A. V., Paul, J., Ward, T. L. (1977). "Discrete time series models of a freeway density process." Proceedings of the 7th International Symposium on Transportation and Traffic Theory, Kyoto, Japan, 387-411.

Gamerman, D. and Migon, H. S. (1993). "Dynamic Hierarchical Models." Journal of the Royal Statistical Society Series B **55**(3): 629-642.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). Bayesian Data Analysis. Boca Raton, Chapman & Hall/CRC.

Gill, P. S. and Swartz, T. B. (2004). "Bayesian analysis of directed graphs data with applications to social networks." Applied Statistics **53**(2): 249-260.

Grunwald, G. K., Hamza, K. and Hyndman, R. J. (1997). "Some properties and generalizations of non-negative Bayesian time series models." Journal of the Royal Statistical Society B **59**(3): 615-626.

Harrison, P. J. (1997). "Convergence and the Dynamic Linear Model." Journal of Forecasting **16**: 287-292.

Hazelton, M. L. (2001). "Estimation of origin-destination trip rates in Leicester." Applied Statistics **50**(4): 423-433.

Heidemann, D. (1999). "Some critical remarks on a class of traffic flow models." Transportation Research Part B **33**(2): 153-155.

Helbing, D., Hennecke, A., Shvetsov, V. and Treiber, M. (2002). "Micro- and macrosimulation of freeway traffic." Mathematical and Computer Modelling **35**(5): 517-547.

Hilliges, M. and Weidlich, W. (1995). "A phenomenological model for dynamic traffic flow in networks." Transportation Research Part B **29**(6): 407-431.

Hjorth, U. (2002). "Traffic subflow estimation and bootstrap analysis from filtered counts." Transportation Research Part B **36**(4): 345-359.

Hodges, J. S. (1998). "Some algebra and geometry for hierarchical models, applied to diagnostics." Journal of the Royal Statistical Society Series B **60**(3): 497-536.

Holland, E. N. (1998). "A generalised stability criterion for motorway traffic." Transportation Research Part B **32**(2): 141-154.

Hoogendoorn, S. P. and Bovy, P. H. L. (2000). "Continuum modeling of multiclass traffic flow." Transportation Research Part B **34**(2): 123-146.

Hoogendoorn, S. P. and Bovy, P. H. L. (2001). "Generic gas-kinetic traffic system modelling with applications to vehicular traffic flow." Transportation Research Part B **35**(4): 317-336.

Hounsell, N. B. and Ishtiaq, S. (1997). "Journey time forecasting for dyhamic route guidance systems in incident conditions." International Journal of Forecasting **13**: 33-42

Huerta, G. and West, M. (1999). "Priors and component structures in autoregressive time series models." Journal of the Royal Statistical Society B **61**(4): 881-899.

Hurdle, V. F and Son, B. (2000). "Road test of a freeway model." Transportation Research Part A **34**(7): 537-564.

Jiang, R., Wu, Q.-S. and Zhu, Z.-J. (2002). "A new continuum model for traffic flow and numerical tests." Transportation Research Part B **36**(5): 405-419.

Jin, W. L. and Zhang, H. M. (2003). "On the distribution schemes for determining flows through a merge." Transportation Research Part B **37**(6): 521-540.

Kirby, H. R., Watson, S. M. et al. (1997). "Should we use neural networks or statistical models for short-term traffic forecasting?" International Journal of Forecasting **13**: 43-50.

Lauritzen, S. L. and Richardson, T. S. (2002). "Chain graph models and their causal interpretations." Journal of the Royal Statistical Society Series B **64**(2): 1-28.

Lighthill, M. J. and Whitham, G. B. (1955). “On kinematic waves I: Flow movement in long rivers II: A theory of traffic flow on long crowded roads.” Proceedings of the Royal Society A229 281-345.

Lin, W.H. and Daganzo, C. F. (1997). "A simple detection scheme for delay-inducing freeway incidents." Transportation Research Part A **31**(2): 141-155.

Makov, U. E. (1983). “Approximate Bayesian Procedures for Dynamic Linear Models in the Presence of Jumps.” The Statistician **32**(1/2): 207-213.

Maunch, M. and Cassidy, M. J. (2002). "Freeway traffic oscillations: observations and predictions." International Symposium of Traffic and Transportation Theory. M. A. P. Taylor. Amsterdam, Elvesier.

McCoy, E. J. and Stephens, D. A. (2004). "Bayesian time series analysis of periodic behaviour and spectral structure." International Journal of Forecasting **20**(4): 713-730.

Mélard, G. and Pasteels, J.-M. (2000). "Automatic ARIMA modelling including interventions, using time series expert software." International Journal of Forecasting **16**(4):497-508

Miller, D. M. and Williams, D. (2004). "Damping seasonal factors: shrinkage estimators for the X-12-ARIMA program." International Journal of Forecasting **20**(4): 529-549.

Nelson, P. (1995). "On deterministic developments of traffic stream models." Transportation Research Part B **29**(4): 297-302.

Nihan, N. L. and Holmesland, K. O. (1980). "Use of the Box and Jenkins Time Series Technique in Traffic Forecasting." Transportation **9**(2): 125-143.

Okutani, I. and Stephanedes, Y. J. (1984). "Dynamic prediction of traffic volume through Kalman filtering theory." Transportation Research Part B **18**(1): 1-11.

Pandher, G. S. (2002). "Forecasting multivariate time series with linear restrictions using constrained structural state-space models." Journal of Forecasting **21**(4): 281-300.

Papageorgiou, M. (1998). "Some remarks on macroscopic traffic flow modelling." Transportation Research Part A **32**(5): 323-329.

Pearl, J. (1996). "Decision making under uncertainty." ACM Computing Surveys **28**(1).

Queen, C. M. (1997). "Model elicitation in competitive markets." The Practice of Bayesian Analysis. S. French and J. Q. Smith, Arnold.

Queen, C. M. and Smith, J. Q. (1993). "Multiregression Dynamic Models." Journal of the Royal Statistical Society Series B **55**(4): 849-870.

Quintana, J. M. and West, M. (1987). "An Analysis of International Exchange Rates Using Multivariate DLM's." The Statistician **36**(2/3): 275-281.

Reilman, M. A., Gunst, R. F. and Lakshminarayanan, M. Y. (1985) "Structural Model Estimation With Correlated Measurement Errors." Biometrika **72**(3): 669-672.

Richards, P. I. (1955). "Shock Waves on the Highway" Operations Research **4**: 42-51.

Rougier, J. and Goldstein, M. (2001). "A Bayesian analysis of fluid flow in pipe-lines." Applied Statistics **50**(1): 77-93.

Roverato, A. and Consonni, G. (2004). "Compatible prior distributions for directed acyclic graph models." Journal of the Royal Statistical Society B **66**(1): 47-61.

Salvador, M., Gallizo, J. L. and Gargallo, P. (2003). "A dynamic principle components analysis based on multivariate matrix normal dynamic linear models." Journal of Forecasting **22**(6-7): 457-478.

Sanwal, K. K., Petty, K., Walrand, J. and Fawaz, Y. (1996). "An extended macroscopic model for traffic flow." Transportation Research Part B **30**(1): 1-9.

Schaalja, G. B. and Butts, R. A. (1993). "Some Effects of Ignoring Correlated Measurement Errors in Straight Line Regression and Prediction." Biometrics **49**(4): 1262-1267.

Schmidt, A. M. and Gamerman, D. (1997). "Temporal Aggregation in Dynamic Linear Models." Journal of Forecasting **16**: 293-310.

Schubert, L. K. (2004). "A new characterisation of probabilities in Bayesian networks." Proceedings of the 20th Conference of Uncertainty in Artificial Intelligence: 495-503.

Smith, B. L. and Demetsky, M. J. (1997). "Traffic flow forecasting: Comparison of modelling approaches." Journal of Transportation Engineering **123**(4): 261-266.

Spiegalhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Statistical Science **8**(3): 219-283.

Tebaldi, C., West, M. and Karr, A. F. (2002). "Statistical Analyses of Freeway Traffic Flows." Journal of Forecasting **21**: 39-68.

Vaage, K. (2000). "Detection of outliers and level shifts in time series: an evaluation of two alternative procedures." Journal of Forecasting **19**(1): 23-37.

van Arem, B., Kirby, H. R., van der Vlist, M. J. M. and Whittaker, J. C. (1997). "Recent advances and applications in the field of short-term traffic forecasting." International Journal of Forecasting **13**: 1-12.

van Arem, B., van der Vlist, M. J. M., et al. (1997). "Travel time estimation in the GERDIEN project." International Journal of Forecasting **13**: 73-85.

van der Zijpp, N. J. and de Romph, E. (1997). "A dynamic traffic forecasting application on the Amsterdam beltway." International Journal of Forecasting **13**: 87-103.

Venables, W. N. and Ripley, B. D. (1999). Modern Applied Statistics with S-Plus, Springer.

Wang, Y. and Papageorgiou, M. (2005). "Real-time traffic state estimation based on extended Kalman filter: a general approach." Transportation Research Part B **39**(2): 141-167.

Watling, D. (1996). "Asymmetric problems and stochastic process models of traffic assignment." Transportation Research Part B **30**(5): 339-357.

Wermuth, N. and Lauritzen, S. L. (1990). "On Substantive Research Hypotheses, Conditional Independence Graphs and Graphical Chain Models." Journal of the Royal Statistical Society Series B **52**(1): 21-50.

West, M. and Harrison, J. (1999). Bayesian Forecasting and Dynamic Models. New York, Springer-Verlag.

Whitlock, M. E. (1999). A Bayesian Approach to Road Traffic Network Modelling. Statistics, Kent: 169.

Whitlock, M. E. and Queen, C. M. (1998). A Bayesian model of a traffic network, technical paper, University of Kent.

Whitlock, M. E. and Queen, C. M. (2000). "Modelling a traffic network with missing data." Journal of Forecasting **19**: 561-574.

Whittaker, J., Garside, S. and Lindveld, K. (1997). "Tracking and predicting a network traffic process." International Journal of Forecasting **13**: 51-61.

Wild, D. (1997). "Short-term forecasting based on a transformation and classification of traffic volume time series." International Journal of Forecasting **13**: 63-72.

Yang, S. and Davis, G. A. (2002). "Bayesian estimation of classified daily traffic." Transportation Research Part A **36**: 365-382.

Yi, J., Lin, H., Alvarez, L. and Horowitz, R. (2003). "Stability of macroscopic traffic flow modeling through wavefront expansion." Transportation Research Part B **37**(7): 661-679.

Zhang, H., Ritchie, S. G. and Lo, Zhen-Ping (1997). "Macroscopic modeling of freeway traffic using an artificial neural network." Transportation Research Record 1588.

Zhang, H. M. (1998). "A theory of nonequilibrium traffic flow." Transportation Research Part B **32**(7): 485-498.

Zhang. H. M. (2001). "A finite difference approximation of a non-equilibrium traffic flow model." Transportation Research Part B **35**(4): 337-365.

Zhang, H. M. (2003). "Driver memory, traffic viscosity and a viscous vehicular traffic flow model." Transportation Research Part B **37**(1): 27-41.

Zhang, H. M. and Kim, T. (2005). "A car-following theory for multiphase vehicular traffic flow." Transportation Research Part B **39**(5): 385-399.

Zio, M. d., Scanu, M., Coppola, L., Luzi, O. and Ponti, A. (2004). "Bayesian networks for imputation." Journal of the Royal Statistical Society Series A **167**(2): 309-322.